EXACT MINIMAX ESTIMATION OF THE PREDICTIVE DENSITY IN SPARSE GAUSSIAN MODELS

By Gourab Mukherjee and Iain M. Johnstone

Stanford University

We consider estimating the predictive density under Kullback-Leibler loss in an ℓ_0 sparse Gaussian sequence model. Explicit expressions of the first order minimax risk along with its exact constant, asymptotically least favorable priors and optimal predictive density estimates are derived. Compared to the sparse recovery results involving point estimation of the normal mean, new decision theoretic phenomena are seen here. Sub-optimal performance of the class of plug-in density estimates reflects the predictive nature of the problem and optimal strategies need diversification of the future risk. We find that minimax optimal strategies lie outside the Gaussian family but can be constructed with threshold predictive density estimates. Novel minimax techniques involving simultaneous calibration of the sparsity adjustment and the risk diversification mechanisms are used to design optimal predictive density estimates.

1. Introduction and main result.

1.1. *Estimating the high-dimensional Gaussian Predictive Density.*

Background: The objective of statistical prediction analysis is to choose a probability distribution which will be good in predicting the behavior of future samples (Aitchison and Dunsmore, 1975). If the observed past data $\mathbf{X}$ and the unobserved future data $\mathbf{Y}$ are generated from a joint density $f(\mathbf{x}, \mathbf{y})$, the objective is to estimate the future conditional density of $f(\mathbf{y}\,|\mathbf{X} = \mathbf{x})$, also referred to as the predictive density (Geisser, 1971). For practical purposes, we usually need to forecast functionals of the predictive density. Good predictive performances can be ensured by using functions of predictive density estimates which are optimally chosen based on appropriate goodness of fit measure.

Here, we consider flexible parametric predictive models outlined in Geisser (1993) which can accommodate most dependencies in the data. We observe $\mathbf{X}$ with $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_{m_1}$ independently generated from a parametric density $f_{(\boldsymbol{\theta}, a_i)}(\cdot)$ indexed by unknown parameters $\boldsymbol{\theta}$ and known parameters

AMS 2000 subject classifications: Primary 62C20; Secondary 62M20, 60G25, 91G70

Keywords and phrases: predictive density, risk diversification, minimax, sparsity, high dimensional, mutual information, plug-in risk, thresholding

$A = \{a_i : 1 \leq i \leq m_1\}$. The future data $\mathbf{Y} = \{\mathbf{Y}_1, \cdots, \mathbf{Y}_{m_2}\}$ are generated from successive independent parametric densities $\{f_{(\boldsymbol{\theta}, b_i)}(\cdot) : 1 \leq i \leq m_2\}$ with invariant (over time) unknown parameters $\boldsymbol{\theta}$ and known parameters $B = \{b_i : 1 \leq i \leq m_2\}$. In such a predictive model the dependence between the past and the future is based on the time-invariant parameters $\boldsymbol{\theta}$.

Generic Parametric Predictive Model: M

$$\begin{array}{lll} \text{PAST} & \text{OBSERVATIONS:} & \mathbf{X}_i \overset{\text{indep.}}{\sim} f_{(\boldsymbol{\theta},\, a_i)}(\cdot)\,, \quad i = 1, \cdots, m_1 \\ \text{FUTURE} & \text{OBSERVATIONS:} & \mathbf{Y}_j \overset{\text{indep.}}{\sim} f_{(\boldsymbol{\theta},\, b_j)}(\cdot)\,, \quad j = 1, \cdots, m_2. \end{array}$$

If $\boldsymbol{\theta}$ is fixed the true predictive density of $\mathbf{Y}$ would be $f_{(\boldsymbol{\theta},B)}(\cdot) = \prod_{j=1}^{m_2} f_{(\boldsymbol{\theta},b_j)}(\cdot)$. We would like to estimate it by density estimates $\widehat{p}(.|\mathbf{X} = \mathbf{x})$. We use the information theoretic measure of Kullback and Leibler (1951) as the goodness of fit measure between the true and estimated distributions

$$\mathbf{L}\left(\boldsymbol{\theta}, \widehat{p}\left(\,\cdot\,|\mathbf{x}\right)\right) = \int f_{(\boldsymbol{\theta},B)}(\mathbf{y})\, \log\left(\frac{f_{(\boldsymbol{\theta},B)}(\mathbf{y})}{\widehat{p}(\mathbf{y}|\mathbf{x})}\right)\, d\mathbf{y}\,.$$

Averaging over the past observations $\mathbf{X}$, the predictive risk of the density estimate $\widehat{p}(\cdot\,|\mathbf{X} = \mathbf{x})$ at $\boldsymbol{\theta}$ is given by

$$(1) \qquad \boldsymbol{\rho}\left(\boldsymbol{\theta}, \widehat{p}\right) = \iint f_{(\boldsymbol{\theta},A)}(\mathbf{x})\, f_{(\boldsymbol{\theta},B)}(\mathbf{y}) \log\left(\frac{f_{(\boldsymbol{\theta},B)}(\mathbf{y})}{\widehat{p}(\mathbf{y}|\mathbf{x})}\right)\, d\mathbf{y}\, d\mathbf{x}\,.$$

The relative entropy predictive risk $\boldsymbol{\rho}\left(\boldsymbol{\theta}, \widehat{p}\right)$ measures the exponential rate of divergence of the joint likelihood ratio over a large number of independent trials (Larimore, 1983). In classical fixed-dimensional parametric analysis, the minimal predictive risk estimate would maximize the expected growth rate in repeated investment scenarios (Cover and Thomas, 1991, Chapter 6 and 15). Competitive optimal predictive schemes (Bell and Cover, 1980) for gambling, sports betting, portfolio selection, etc can be constructed from predictive density estimates with optimal Kullback-Leibler (KL) risk properties. In data compression set-up $\mathbf{L}\left(\boldsymbol{\theta}, \widehat{p}\left(\,\cdot\,|\mathbf{x}\right)\right)$ reflects the excess average code length that we need if we use the conditional density estimate $\widehat{p}$ instead of the true density to construct a uniquely decodable code for the data $\mathbf{Y}$ given the past $\mathbf{x}$ (McMillan, 1956). The notion can be extended to a sequential framework where minimizing the predictive risk would result in the minimum description length (Barron, Rissanen and Yu, 1998, Rissanen, 1984) based estimate of the true parametric density (Liang and Barron, 2005).

The modern data deluge has influenced a rapid evolution of statistical methods towards simultaneous multi-parametric analyses (Donoho, 2000). The traditional fixed dimensional predictive density estimation problem needs extension to high-dimensional parametric models in the following applications:

Data compression: Coding of high-dimensional data (Candès, 2006, Liang and Barron, 2004) needs construction of a decodable code for $\mathbf{Y}$ given the value of $\mathbf{X}$. If the high-dimensional parameter $\boldsymbol{\theta}$ is known the optimal expected length of such a code would have been based on the true density $f_{(\boldsymbol{\theta},B)}$. In universal data compression (Rissanen, 1984), without any prior knowledge of $\boldsymbol{\theta}$, a choice of predictive density $\widehat{p}\,(\mathbf{Y}|\mathbf{x})$ will be used instead of the true density to construct the code. The excess average code length in that case is given by:

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\log_2(1/\widehat{p}(\mathbf{Y}|\mathbf{x})) - \log_2(1/f_{(\boldsymbol{\theta},B)}(\mathbf{Y}))\right] = \mathbf{L}\big(\boldsymbol{\theta}, \widehat{p}\,(.|\mathbf{x})\big)/\log 2 \text{ bits}. \tag{2}$$

We ignore the issue of discretization as the loss will be the limit redundancy based on infinitesimally fine choice of discretizations (Csiszár, 1973, Vajda, 2002). Now, if the parameter $\boldsymbol{\theta}$ was generated from the distribution π then the minimal excess average code length is given by the Bayes risk of the prior π on $\boldsymbol{\theta}$ (Liang, 2002). The integrated Bayes risk of the estimate $\widehat{p}$ for the prior π is given by $B(\pi, \widehat{p}) = \int \pi(\boldsymbol{\theta})\boldsymbol{\rho}\left(\boldsymbol{\theta}, \widehat{p}\right) d\boldsymbol{\theta}$ and the Bayes risk $B(\pi)$ of the prior equals $\min_{\widehat{p}} B(\pi, \widehat{p})$. It is equal to the mutual conditional information $I_\pi(\Theta; \mathbf{Y}|\mathbf{X})$ between the unknown parameter $\boldsymbol{\theta}$ and the future data $\mathbf{Y}$ given the past $\mathbf{X}$ (see Appendix section A.1 for the results in the case of Gaussian channels). Intuitively, based on the decomposition of $I_\pi(\Theta; \mathbf{Y}|\mathbf{X}) = I_\pi(\Theta; \mathbf{X}, \mathbf{Y}) - I_\pi(\Theta; \mathbf{X})$ it can be seen that minimizing the Bayes predictive risk $B(\pi, \widehat{p})$ would signify extracting the maximum possible dependence of $\mathbf{Y}$ and Θ based on $\mathbf{X}$. The information capacity of the channel C is given by the maximal mutual conditional information $\max_{\pi\in\mathcal{M}} I_\pi(\Theta; \mathbf{Y}|\mathbf{X})$ over an appropriate class of priors. Here, we will evaluate C by explicitly calculating the maximal Bayes risk $\max_{\pi\in\mathcal{M}} B(\pi)$.

Sequential Investment with side information: Investment schemes based on high-frequency trading need predictive strategies on financial instruments governed by a large number of parameters (Fan, Lv and Qi, 2011). The log-optimal predictive strategies of Barron and Cover (1988) will depend on the high-dimensional density estimates minimizing the predictive risk in Equation [1].

Sports Betting: Online betting portals have not only increased traffic but also caused a massive transformation of the fixed-odds sports betting market (Buchdahl, 2003). Betfair [1], one of the leading betting exchanges in U.K. matches 15 times as many daily transactions as the London Stock Exchange. These online stochastic markets allow bets with the most lucrative odds to be placed on the joint occurrences of several events (multiple bets). As historical data can be accessed through portal-supplied application programming interface, statistical techniques are being increasingly used in designing betting strategies (Magee, 2011) and multi-parametric models are required to estimate the multiple-bets probabilities.

Often, in high dimensional models transformations of the data approximate to normality are available. So henceforth we impose Gaussian distribution assumptions and in particular consider prediction in the location model:

High Dimensional Gaussian Predictive Model

$$\textbf{M.1} \qquad \mathbf{X} \sim N(A\,\boldsymbol{\theta},\, \sigma_p^2\, I) \quad \text{and} \quad \mathbf{Y} \sim N(B\,\boldsymbol{\theta},\, \sigma_f^2\, I)$$

where A and B are $m_1 \times n$ and $m_2 \times n$ data matrices with n being very large, σ_p^2 and σ_f^2 are the volatilities of the past $\mathbf{X}$ and the future $\mathbf{Y}$. The location structure depend on the time-invariant unknown vector $\boldsymbol{\theta}$ of length n. When $\boldsymbol{\theta}$ is dense (non-sparse), predictive density estimation in **M.1** has been discussed in Mukherjee and Johnstone (2012b). Here, we further assume that $\boldsymbol{\theta}$ is sparse with few non-zero coefficients. We impose an $\boldsymbol{\ell_0}$ constraint on the parameter space:

$$\Theta(n, s) = \left\{ \boldsymbol{\theta} \in \mathbb{R}^n : \sum_{i=1}^{n} \mathbb{I}[\theta_i \neq 0] \leq s \right\}. \tag{3}$$

This notion of sparsity is widely used in modeling highly interactive systems (represented by a large number of related parameters) which are dominated by only few significant effects. Sparse models have been successfully employed in biological sciences (Tibshirani et al., 2002), engineering applications (Donoho, 2006) and financial modeling (Brodiea et al., 2009). The predictive model **M.1** with $\boldsymbol{\ell_0}$ constraint on the location structure can be used for sparse coding and for prediction in sparse networks.

Here, the minimax risk calculations will be based on the orthogonal Gaussian model. Like sparse point estimation (Raskutti, Wainwright and Yu, 2011, Zhang, 2010), constrained predictive density estimation in **M.1** would

[1] http://sports.betfair.com/

intrinsically depend on risk calculations in the orthogonal model:

Predictive Gaussian Sequence Model

$$\mathbf{M.2} \qquad \mathbf{X} \sim N(\boldsymbol{\theta},\, \sigma_p^2\, I) \quad \text{and} \quad \mathbf{Y} \sim N(\boldsymbol{\theta},\, \sigma_f^2\, I)$$

where $\mathbf{X}$ and $\mathbf{Y}$ are both n – dimensional vectors. **M.2** is known as the homoscedastic Gaussian sequence model (Nussbaum, 1996) and has been widely studied in the function estimation framework (Johnstone, 2012). Optimal estimation in **M.1** can be linked with the minimax decision theoretic results in **M.2** through the procedure outlined in Donoho, Johnstone and Montanari (2011).

Sparse Gaussian predictive density estimation has the attributes of a sparse prediction problem adapted to the peculiarities of the entropy loss function. Point prediction analyses of dense signals in **M.1** (Dicker, 2012, Huber and Leeb, 2012, Leeb, 2009) relate the worst-case performance with the spectral distribution of the predictors. Here, we concentrate on the orthogonal model. We address some of the unresolved issues associated with the role of sparsity in prediction theory. Next, we describe in more detail the minimax predictive density estimation problem in the sparse orthogonal model **M.2**.

1.2. *Minimax estimation under sparsity constraints.*

In order to help the reader to better understand the context and nature of the predictive problem, we provide a brief review of the literature around the predictive density estimation problem. Aitchison (1975), Murray (1977), Ng (1980) showed that in most parametric models there exist Bayes predictive density estimates which are decision theoretically better than the maximum likelihood plug-in estimate. As the name suggests, a plug-in or estimative density estimate $f_{(\widehat{\boldsymbol{\theta}},\, B)}$ belongs to the same parametric family of the true density and has the point estimate $\widehat{\boldsymbol{\theta}}$ plugged in the place of the unknown parameter. Given a prior π over $\mathbb{R}^n$ the Bayes predictive density in $\mathbf{M}$ (along with some mild conditions) minimizes the intergrated Bayes risk and is given by

$$(4) \quad \widehat{p}_{\,\pi}\left(\mathbf{y}|\mathbf{X}=\mathbf{x}\right) = \int f_{(\boldsymbol{\theta},B)}(\mathbf{y})\, \pi(\boldsymbol{\theta}|\,\mathbf{x})\, d\boldsymbol{\theta}$$

where the posterior distribution

$$(5) \quad \pi(\boldsymbol{\theta}\,|\mathbf{x}) = \{m_\pi(\mathbf{x})\}^{-1}\, f_{(\boldsymbol{\theta},A)}(\mathbf{x})\, \pi(\boldsymbol{\theta}) \text{ and } m_\pi(\mathbf{x}) = \int f_{(\boldsymbol{\theta},A)}(\mathbf{x})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

is the marginal distribution. An important issue in predictive inference has always been to compare the performance of the class $\mathcal{E}$ of point estimation (PE) based plug-in density estimates (Barndorff-Nielsen and Cox, 1996) with that of the optimal predictive density estimate. In fixed dimensional parametric spaces, large sample attributes of the predictive risk of efficient plug-in and Bayes density estimates have been studied by Komaki (1996) and Hartigan (1998). These results are independent of specific distributional attributes of f (Aslan, 2006) and reflect the predictive nature of the problem through the relative inefficiency of the maximum likelihood plug-in density estimates.

Recently, the predictive density estimation problem has been studied in high dimensional parametric spaces (George, Liang and Xu, 2006, Ghosh, Mergel and Datta, 2008, Komaki, 2004, Xu and Zhou, 2011). Decision theoretic parallels between predictive density estimation under Kullback-Leibler loss and point estimation under quadratic loss have been explored in **M.2** (George, Liang and Xu, 2012). Fundamental techniques and results in unconstrained Gaussian point estimation theory (Brown, 1971, Brown and Hwang, 1982, Stein, 1974, Strawderman, 1971) can be extended to produce optimal predictive density estimates (Brown, George and Xu, 2008, Fourdrinier et al., 2011, Komaki, 2001). Table 1 summaries these decision theory results in the point estimation and predictive density estimation regimes. The Bayes predictive density from the uniform prior $\widehat{p}_U$ is the best invariant as well as a minimax density estimate in the unconstrained parametric space. Its risk properties are similar to those of the canonical minimax point estimate $\mathbf{X}$. Both regimes exhibit inadmissibility of the best invariant estimates in their respective domains and improved minimax estimators are constructed.

Another important subclass of density estimates are Linear estimates ($\mathcal{L}$) which are Bayes rules based on the conjugate product normal priors. The resultant density estimates $\widehat{p}_L[\,\boldsymbol{\alpha}\,] = \prod_{i=1}^n N(\alpha_i X_i, \alpha_i + \sigma_f^2)$, with $\alpha_i \in [0,1]$, are still Gaussian but has larger variance than the future density $f_{\boldsymbol{\theta},\sigma_f^2}(\mathbf{y})$. We choose the name 'linear' because the conjugate prior implies linearity of the posterior mean in X. It should be noted that for linear estimates, shrinkage of the location estimate $\mathbf{X}$ is related to flattening of the variance and $\mathcal{E} \cap \mathcal{L}$ consists only the zero density estimate.

Xu and Liang (2010) showed that the class $\mathcal{L}$ is minimax optimal if the parameter space is restricted to ellipsoids with certain growth conditions. Here, we evaluate the minimax risk over the $\boldsymbol{\ell}_0$ sparse parameter space $\Theta(n,s)$ in the asymptotic framework $\{n \to \infty$ and $s/n \to 0\}$. Sparse point estimation has been extensively studied in this asymptotic set-up by Donoho and Johnstone (1994a,b), Donoho et al. (1992), Foster and George (1994) and the results

are the building blocks for popular sparse estimation methods (Candes and Tao, 2007, Donoho, Maleki and Montanari, 2011, Zhang, 2005). It is natural to look for parallels in the predictive density regime.

TABLE 1
Parallels between Point Estimation of the Normal mean under quadratic loss and Predictive Density Estimation under KL loss

Decision Theoretic Issues	Point Estimation	Predictive Density
Admissibility in Unrestricted space		
Shrinkage priors	Stein (1974) Strawderman (1971)	Komaki (2001) George, Liang and Xu (2006)
Complete Class & Bayes Rules	Brown and Hwang (1982)	Brown, George and Xu (2008)
Minimaxity over & Restricted space		
Ellipsoids	Pinsker (1980)	Xu and Liang (2010)
Sparsity Constraints	Donoho et al. (1992)	Evaluated here

Our contribution: Instead of parallels, we found contrasting results for sparse estimation in the two regimes. The asymptotic minimax predictive risk reflects the nature of the predictive density estimation problem through the ratio r of the future to the past volatilities $r = \sigma_f^2/\sigma_p^2$. As r decreases, we need to estimate the future observations based on increasingly noisy past observations and so, the difficulty of the density estimation problem also increases. Unlike point estimation, sharp decision theoretic rates in the predictive density problem should depend on r. This dependence was not emphasized in the admissibility results in the unrestricted space.

In our ℓ_0 sparse prediction framework, as the proportion of non-zero signals goes to zero, we find that the order of the minimax rate does not depend on r. So, exact determination of the constants of the minimax risk is important here. Optimal minimax estimators can be constructed by incorporating the predictive nature of the problem through the notion of *diversification* of the future risk. Under sparsity constraints efficiency of the prediction schemes depend on careful coupling of the sparsity adjustment and the risk diversification mechanisms. The risk diversification notion can also be extended (though not done here) to dense unrestricted parametric spaces where

future uncertainty can be effectively shared by optimally flattening probability densities based on the quadratic risk estimate of their corresponding location point estimator.

Here we also evaluate the minimax risk over the wide class $\mathcal{G}$ of all product *Gaussian density estimates* $\widehat{p}_G[\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{d}}] = \prod_{i=1}^n N(\widehat{\theta}_i, \widehat{d}_i)$. $\mathcal{G}$ contains both $\mathcal{L}$ and $\mathcal{E}$ and would represent the infamily error rate of the sparse Gaussian predictive density estimation problem. We prove that the sub-class $\mathcal{G}$ is sub-optimal rand provide asymptotic miximax strategies as well as the sub-optimality rates of the sub-classes $\mathcal{E}, \mathcal{L}$ and $\mathcal{G}$.

1.3. *Description of the main results.*

Notations and Preliminaries. To proceed further we need some notation. The action space $\mathcal{A}_n$ contains all possible densities in $\mathbb{R}^n$. The n-dimensional minimax risk of the prediction problem is given by

$$R(n, s, r) = \min_{\widehat{p} \in \mathcal{A}_n} \max_{\boldsymbol{\theta} \in \Theta(n,s)} \rho\left(\boldsymbol{\theta}, \widehat{p}\right).$$

We compute the limiting behaviour of $R(n, s, r)$ in the asymptotic framework $\mathcal{F} = \{n \to \infty, s/n \to 0\}$. The minimax risk over the sub-class of Plugin $(\mathcal{E})$ density estimates is represented by

$$R(n, s, r, \mathcal{E}) = \min_{\widehat{\boldsymbol{\theta}}} \max_{\boldsymbol{\theta} \in \Theta(n,s)} \rho\left(\boldsymbol{\theta}, \widehat{p}_E[\widehat{\boldsymbol{\theta}}]\right) \text{ where } \widehat{p}_E[\widehat{\boldsymbol{\theta}}] = N(\widehat{\boldsymbol{\theta}}, \sigma_f^2 \mathrm{I}_n).$$

Similarly, the minimax risk over the sub-classes of Linear $(\mathcal{L})$ and Gaussian density estimates $(\mathcal{G})$ will be denoted by $R(n, s, r, \mathcal{L})$ and $R(n, s, r, \mathcal{G})$ respectively. The maximum Bayes risk over the class of priors $\mathcal{M}(n)$ on $\mathbb{R}^n$ is denoted by

$$B(r, \mathcal{M}(n)) = \max_{\pi \in \mathcal{M}(n)} \min_{\widehat{p}} B(\pi, \widehat{p}).$$

As defined in Subsection 1.1 this maximin value is also the information capacity. A prior maximizing this Bayes risk is said to be a least favorable prior for the prediction problem. We evaluate the supremum Bayes risk of the following class of priors

$$\mathcal{M}(n, s) = \left\{ \pi : \sum_{i=1}^n P_\pi(\theta_i \neq 0) \leq s \right\}.$$

Univariate Prediction Problem. In high dimensions, due to concentration of measure, the decision theoretic results in our multivariate set up $\mathcal{F}$ will intrinsically depend on the properties of the coordinate-wise univariate risk.

The least favorable priors as well as the minimax density estimates will be product densities. So, computing the multivariate risk in the n–dimensional, s–sparse orthogonal Gaussian Model $\mathbf{M.2}(n, s, \sigma_p, \sigma_f)$ would involve studying the corresponding univariate model $\mathbf{M.2}(1, \eta, \sigma_p, \sigma_f)$ in which we relax the sparsity constraint to restriction on the univariate prior space

$$\mathfrak{m}(\eta) = \{\pi \in \mathcal{P}(\mathbb{R}) : \pi(\theta \neq 0) \leq \eta\}$$

where $\mathcal{P}(\mathbb{R})$ is the collection of all probability measures in $\mathbb{R}$. The maximin value $\sup_{\pi\in\mathfrak{m}(\eta)} \inf_{\widehat{p}} B(\pi, \widehat{p})$ of this univariate prediction game is given by the maximal Bayes risk $\beta(\eta, r) := \sup_{\pi\in\mathfrak{m}(\eta)} B(\pi)$. The minimax risk for this univariate prediction problem is given by

$$\rho_M(\eta, r) := \inf_{\widehat{p}} \sup_{\pi\in\mathfrak{m}(\eta)} B(\pi, \widehat{p}). \tag{6}$$

The minimax risk and the maximal Bayes risk over univariate sub-collection $\mathfrak{m}$ of priors of $\mathfrak{m}(\eta)$ are respectively denoted by $\rho_M(\eta, r, \mathfrak{m})$ and $\beta(\eta, r, \mathfrak{m})$. When the maximal Bayes risk (maximin) equals to the minimax risk, it is referred as the Bayes-Minimax risk for the prediction problem.

As in our asymptotic framework $\mathcal{F}$ the proportion of non-zero signals s/n goes to zero, the univariate risk calculation will be in the asymptotic regime $\eta \to 0$. The difference between the multivariate and univariate cases is notationally demonstrated through the bold representation of multivariate vectors. The other non-standard notations used are $\phi(|\boldsymbol{\theta}, r)$ for the multivariate normal density with center $\boldsymbol{\theta}$ and covariance $r\mathrm{I}$ while $\widetilde{\Phi} = 1 - \Phi$ with Φ being the standard normal distribution. For sequences, the symbol $a_n \sim b_n$ means $a_n = b_n(1 + \mathrm{o}(1))$ and $a_n \asymp b_n$ means $a_n/b_n \in (c_1, c_2)$ where c_1 and c_2 are constants.

Results. Consider the following symmetric univariate prior 3–point prior

$$\pi[\eta, r, 3] = (1 - \eta) \cdot \delta_0 + \frac{1}{2}\,\eta \cdot \delta_{\nu_\eta} + \frac{1}{2}\,\eta \cdot \delta_{-\nu_\eta}$$

where ν_η is the positive root of the quadratic equation

$$v_w^{-1}\,\nu^2 + 2\,v_w^{-1/2}\,\nu\,a = \lambda_e^2$$

with $v_w = (1+r^{-1})^{-1}$, $a = \max(\sqrt{\log v_w\lambda_e^2}, 1)$ and $\lambda_e^2 = 2\sigma_p^2 \log\{(1-\eta)\,\eta^{-1}\}$ is close to the universal threshold (when $\eta = n^{-1}$) seen previously in PE (Donoho and Johnstone, 1994a). As $\eta \to 0$, the solution $\nu_\eta \to \lambda_f = v_w^{-1/2}\lambda_e$.

Also, consider the discrete cluster prior $\pi[\eta, r, \mathrm{CL}]$ with $1-\eta$ probability at 0 and sharing the remaining mass among a cluster of support points. The non-zero support points start from $\pm\,\nu_\eta$ and span out symmetrically on either side in a geometric progression with common ratio $(1+2r)$ up to the universal threshold:

$$\pi[\eta, r, \mathrm{CL}] = (1-\eta)\cdot\delta_0 + \frac{\eta}{2K_\eta}\sum_{i=1}^{K_\eta}\left\{\,\delta_{\mu_i} \,+\, \delta_{-\mu_i}\,\right\} \quad \text{where} \tag{7}$$

$$K_\eta = \max\left\{i : (1+2r)^{\,i-1}\,\nu_\eta \le \lambda_e + a\right\}, \tag{8}$$

$$\mu_i = (1+2\,r)^{i-1}\;\nu_\eta,\; i = 1, 2, \ldots, K_\eta. \tag{9}$$

For any fixed $r \in (0,\infty)$ as $\eta \to 0$ we have

$$K(r) = \lim_{\eta\to 0} K_\eta = \left\lfloor \frac{\log(1+r^{-1})}{2\log(1+2r)} \right\rfloor.$$

We will use the Bayes predictive density $\widehat{p}\,(\cdot|x;\pi[\eta, r, \mathrm{CL}])$ derived from the cluster prior $\pi[\eta, r, \mathrm{CL}]$ to construct threshold estimates. Consider the following univariate threshold estimate which uses the best invariant density estimate $\widehat{p}\,(\cdot|x;\pi_U)$ from the uniform prior above the threshold λ_e and $\widehat{p}\,(\cdot|x;\pi[\eta, r, \mathrm{CL}])$ below the threshold

$$\widehat{p}[\eta, \mathrm{T}, \mathrm{CL}, \mathrm{U}](y|x) = \begin{cases} \widehat{p}\,(y|x;\pi[\eta, r, \mathrm{CL}]) & \text{if } |x| \le \lambda_e \\ \widehat{p}\,(y|x;\pi_{\mathrm{U}}) & \text{if } |x| \ge \lambda_e \end{cases}. \tag{10}$$

The estimator $\widehat{p}[\eta, \mathrm{T}, \mathrm{CL}, \mathrm{U}]$ attains the univariate minimax risk as $\eta \to 0$.

THEOREM 1.1. *For any fixed $r \in (0,\infty)$ as $\eta \to 0$ in the univariate prediction problem we have*

$$\rho_M(\eta, r) = \beta(\eta, r) = \frac{\eta\, v_w\, \lambda_e^2}{2r}\Big(1 + o\,(1)\Big).$$

Also, $\pi[\eta, r, 3]$ is an asymptotically least favorable prior and $\widehat{p}[\eta,$ T, CL, U$]$ is an asymptotically minimax optimal estimate.

Based on the univariate version, we can construct a multivariate, coordinate wise rule

$$\widehat{p}[n, s, \mathrm{T}, \mathrm{CL}, \mathrm{U}](\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n}\widehat{p}\,[s/n, \mathrm{T}, \mathrm{CL}, \mathrm{U}](y_i|x_i)$$

which will be asymptotically minimax optimal in the high dimensional regime $\mathcal{F}$. Also the product discrete distribution

$$\pi[n, s, r, 3](\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi[s/n, r, 3](\theta_i)$$

based on the 3–point prior will be asymptotically least favorable. The following theorem, which is *our main result*, describes the minimaxity results for the predictive density estimation problem with ℓ_0 sparsity constraints in Model **M.2**.

THEOREM 1.2. *As $n \to \infty$, if $s \to \infty$ but $s/n \to 0$ then for any fixed $r \in (0, \infty]$ we have:*

a. *The minimax risk $R(n, s, r) \sim (1 + r)^{-1}\, s \log(n/s)$.*

b. *$\pi[n, s, r, 3]$ is an asymptotically least favorable prior distribution, i.e.*

$$B(\pi[n, s, r, 3]) \sim \sup_{\pi \in \mathcal{P}(\Theta(n,s))} \inf_{\widehat{p} \in \mathcal{A}_n} B(\pi, \widehat{p})$$

where $\mathcal{P}\big(\Theta(n, s)\big)$ is the collection of all probability measures over $\Theta(n, s)$.

c. *The predictive density estimate $\widehat{p}[n, s, T, CL, U]$ is minimax optimal, i.e.*

$$\max_{\boldsymbol{\theta} \in \Theta(n,s)} \boldsymbol{\rho}\left(\boldsymbol{\theta}, \widehat{p}\,[n, s, T, CL, U]\right) \sim R\,(n, s, r).$$

We compute the multivariate minimax risk over the different sub-classes of predictive density estimates. As an immediate corollary of the above theorem it follows that the class of plug-in estimators $\mathcal{E}$ is sub-optimal. The plug-in sup-optimality ratio $R(n, s, r, \mathcal{E})/R(n, s, r)$ asymptotically equals $1 + r^{-1}$ (see Lemma 6.1). As in point estimation, the class of linear estimates $\mathcal{L}$ performs very poorly.

LEMMA 1.1. *For any fixed $r \in (0, \infty)$ and for all sequences s_n such that $s_n \to \infty$ and $s_n/n \to 0$ as $n \to \infty$, we have*

$$\liminf_{n \to \infty} \quad \frac{R(n, s_n, r, \mathcal{L})}{R(n, s_n, r)} = \infty\,.$$

We also find that the performance of the wider class of all Gaussian density estimates is no better than that of plug-in estimates.

LEMMA 1.2. *Under the condition of Theorem 1.2 we have*

$$R\big(n, s, r, \mathcal{G}\big) \sim R\big(n, s, r, \mathcal{E}\big).$$

If the parametric space $\Theta(n, s)$ does not have any sparser representation with respect to the group of orthogonal transformations then the sub-optimality of the class $\widetilde{\mathcal{G}}$ comprising of all Gaussian densities $N(\boldsymbol{\theta}, \widehat{\Sigma})$ including non-diagonal covariances is $1 + r^{-1}$.

Now, $\widehat{p}\,[n, s, \mathrm{T}, \mathrm{CL}, \mathrm{U}]$ is not derived from a prior and we would like to construct a prior for which asymptotic minimaxity in Theorem 1.2 holds. Consider another symmetric univariate prior $\pi[\eta, r, \mathrm{INF}]$ whose support consists of the origin and infinite number of equidistant clusters each containing $2K_\eta$ points (defined before in equation [7]) in the same spatial alignment as for $\pi[\eta, r, \mathrm{CL}]$. As $\eta \to 0$, the clusters centers are separated by λ_e and, as they move away from zero they have geometrically decaying probability with η being the common ratio. However, within clusters all support points are not equally likely any more. They have geometrically decaying probability with common ratio $\log \eta^{-1}$.

$$\pi[\eta, r, \mathrm{INF}] = (1-\eta)\cdot\delta_0 + \frac{1-\eta}{2}\sum_{j=0}^{\infty}\eta^{j+1}\sum_{i=1}^{K_\eta} q_i\left[\delta_{\mu_{ij}} + \delta_{-\mu_{ij}}\right] \quad \text{where,}$$

$$\mu_{ij} = j\,\lambda_e + (1+2r)^{i-1}\,\nu_\eta\,,\ i = 1, \ldots, K_\eta \text{ and } j = 1, \ldots, \infty;$$

$$q_i = (\log \eta^{-1})^{-i} \text{ for } i = 2, \ldots, K_\eta \text{ and } q_1 = 1 - \frac{1 - (\log \eta^{-1})^{-K_\eta}}{\log \eta^{-1} - 1}.$$

Based on $\pi[\eta, r, \mathrm{INF}]$ we can construct a multivariate prior $\pi[n, s, r, \mathrm{INF}](\boldsymbol{\theta}) = \prod_{i=1}^{n}\pi[s/n, r, \mathrm{INF}](\theta_i)$ in $\mathbb{R}^n$ which will not only be least favorable but also yield a minimax optimal density estimate. As $\pi[n, s, r, \mathrm{INF}]$ is a proper prior it is admissible. Though its support is not confined to $\Theta(n, s)$ it concentrates on it asymptotically. It represents an equilibrium solution for the sparse minimax prediction problem.

THEOREM 1.3. *Under the conditions of Theorem 1.2 for any fixed $r \in (0, \infty]$, the proper prior distribution $\pi[n, s, r, INF]$ is an asymptotically least favorable and its corresponding Bayes predictive density is asymptotically minimax optimal.*

These results reflect the predictive nature of the problem. The cluster prior $\pi[\eta, r, \mathrm{CL}]$ in the minimax estimate $\widehat{p}[\eta, \mathrm{T}, \mathrm{CL}, \mathrm{U}](y|x)$ diversifies the predictive risk over the constrained parametric space. The risk diversification

notion is essential to construction of optimal estimates and can be extended to other different forms of asymptotically minimax predictive density estimates. This mechanism of uncertainty sharing in presence of sparsity has not been previously described in minimax decision theory. To rigorously interpret the results, we need the risk equations in George, Liang and Xu (2006) which connect the Bayes predictive risk and with the square error of the posterior mean (see Section 2.1). Next, we provide an heuristic explanation of the implications of the results by an asymptotic (as $\eta \to 0$) risk analysis of univariate threshold estimators.

New Phenomena in Estimation Theory. To adjust for high sparsity we use threshold based non-linear estimates $\widehat{t}[\lambda, S]$ with the threshold cut off at λ, the best invariant estimate $\widehat{p}(\cdot|x;\pi_U)$ above the threshold and estimate/scheme S below the threshold. We found that for such an estimate the threshold choice is dictated by the level of sparsity η and can not be lower than λ_e.

LEMMA 1.3. *For any fixed $r \in (0,\infty)$, scheme S and $u \in [0,1)$, we have*

$$\lim_{\eta\to 0} \frac{\sup_{\pi\in\mathfrak{m}(\eta)} B\big(\pi, \widehat{t}\,[u\,\lambda_e,\, S]\,\big)}{\rho_M(\eta, r)} = \infty.$$

However, unlike PE, here the non-zero support point of the least favorable prior is not at λ_e but at λ_f. So, the univariate asymptotic maximal Bayes risk $\beta(\eta, r) \sim (2\,r)^{-1}\eta\,\lambda_f^2$ is lower than the corresponding maximal quadratic Bayes risk $\beta_q(\eta, r) = \eta\,\lambda_e^2$ after adjustment by $2r$. Because of the threshold choice, the univariate threshold risk $\rho(\theta, \widehat{t}[\lambda_e, \cdot])$ is bounded when $|\theta| \geq \lambda_e$. So, we need to restrict the predictive univariate risk in the region $\{\theta \in (-\lambda_e, \lambda_e)\}$ below the minimax risk. For that purpose, unlike PE we can not just use the zero estimator $\phi(\cdot|0, \sigma_f^2)$ below the threshold because then $\rho(\theta, \widehat{t}[\lambda_e, 0])$ exceeds $\eta^{-1}\,\rho_M(\eta, r)$ (given by equation [6]) in the region $V = \{\theta : |\theta| \in [\lambda_f, \lambda_e]\}$. It leads to inefficiency of the optimal plug-in density estimates.

Instead using the Bayes density estimate from $\pi[\eta, r, 3]$ the risk can be controlled in the neighborhood around λ_f but exceeds $\beta(\eta, r)$ as θ moves further away. The univariate threshold estimate $\widehat{t}[\lambda_e, \pi[\eta, r, 3]]$ represents an unshared threshold prediction scheme. To control the risk through out we need to share the predictive risk for θ between λ_f and λ_e. The cluster prior serves the purpose by using a prior with probability $1 - \eta$ at 0 (which controls the risk at 0) and distributing the remaining mass η equally among a finite chain of points covering V. We also find that discreteness of the

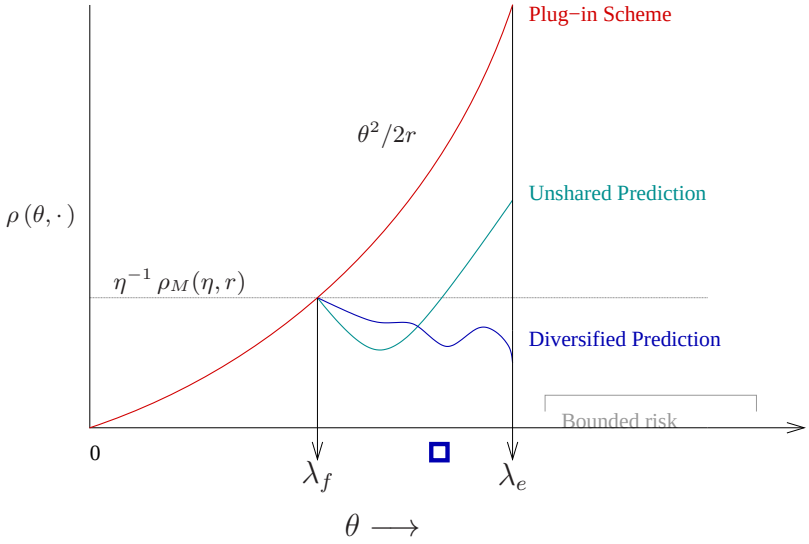

FIG 1. *Schematic diagram of KL-risk functions for different Predictive Schemes. As the true parameter θ varies, the univariate asymptotic predictive risk $\lim_{\eta\to 0}\,\rho(\theta,\widehat{t}[\lambda_e,S])$ is represented on the ordinate. The blue box between λ_f and λ_e represents a support point of the cluster prior (representative of shared predictive schemes) which is not in the support of $\pi[\eta,r,3]$ or other unshared predictive schemes.*

sharing scheme is important and continuous uniform sharing scheme will not work here. Also, the number of support points in the sharing scheme is proportional to r^{-1} reflecting the increasing difficulty of the prediction problem. Table 2 shows the number of support points in the cluster prior as r varies. In Figure [1], we have a schematic description of the asymptotic $(\eta\to 0)$ behavior of $\rho(\theta,\widehat{t}[\lambda_e,S])$ for different type of schemes in S.

TABLE 2

Number (K_η) of positive support points in the cluster prior $\pi[\eta,r,CL]$ as r varies.

r	0.1073	0.1235	0.1465	0.1826	0.2485	0.4196	>0.4196
K_η	7	6	5	4	3	2	1

1.4. *Organization of the paper.* The proof of the results along with their implications are developed first in an overview fashion in Section 2 which may suffice for a first reading. Along with the general proof strategies it contains the proof of Theorem 1.1. Part of the proof of Theorem 1.2 extends over to Section 5. Technical proofs of all the statements in Section 2 are presented in Section 3 and Section 4 with some of the lemmas pushed to the Appendix to improve flow and readability. Theorem 1.3, Lemma 1.1,

Lemma 1.2 and Lemma 1.3 are proved in Section 6. The proofs involving direct risk calculations with the KL loss may be of independent interest and use in information theory.

2. Proof overview and interpretation of the results.

Hereon we will assume that $\sigma_p^2 = 1$ and $\sigma_f^2 = r$. The general predictive KL risk as well as the ℓ_0 constraints on the parametric space will not be affected by this restriction. However, the density estimates are usually based on statistics equivariant to the scale transformation and needs multiplication by σ_p. The proofs as well as the interpretation which will be mostly done on the reduced univariate model are presented for the case $\sigma_p^2 = 1$ and $\sigma_f^2 = r$. While extending the results to the multivariate set-up we will appropriately modify the estimators for general σ_p and σ_f. Proper interpretation of the predictive results will involve comparison with quadratic risk of point estimation in model $\mathbf{M.2}(1, \eta, r)$. Next, we describe the connections for the univariate version.

2.1. *Relations with Point Estimation: Connecting Equations and path of experiments.*

The Kullback-Leibler (KL) predictive risk is connected to risk calculations in point estimation (PE) theory via the semi-futuristic random variable $W = v_w\,(X + r^{-1}Y)$ where $v_w = (1 + r^{-1})^{-1}$. W would have been the UMVUE for the unknown location parameter θ if the future Y were also known along with the past X. The connections between the predictive and point estimation theory center around the parallel to the Tweedie's formula (Brown, 1971, Efron, 2011, Robbins, 1956) which gives a cloased from expression of the Bayes estimate $\widehat{\theta}_\pi$ corresponding to prior π for the location θ estimation problem under quadratic loss

$$\widehat{\theta}_\pi\big(X\big) = X + \nabla \log m_\pi(X, 1) \tag{11}$$

where $m_\pi(Z, v) = \int \phi(Z \,|\, \theta, v)\, \pi(\theta)\, d\theta$ denote the marginal distribution of a Gaussian random variable Z with variance v and with prior distribution π on the location parameter θ. Bayes predictive densities for KL loss in $\mathbf{M.2}(1, \eta, r)$ is analogously related to the best invariant density estimate $\widehat{p}_{\mathrm{U}}$:

$$\widehat{p}_\pi\big(Y \big|\, X = x\big) = \big\{m_\pi\big(W_x, v_w\big)\, m_\pi^{-1}\big(x, 1\big)\big\} \times \widehat{p}_{\mathrm{U}}\big(Y \big|\, X = x\big) \tag{12}$$

where $W_x = v_w\,(x + r^{-1}Y)$. As such, the Bayes risk in these two regimes are also related. By Brown, George and Xu (2008, Theorem 1) the predictive

risk of any prior π in $\mathbf{M.2}(1,\eta,r)$ with $\sigma_p^2 = 1$ is given by $m_\pi(z;1) < \infty$ for all $z \in \mathbb{R}$ we have,

$$\rho(\theta, \widehat{p}_\pi) = \frac{1}{2}\int_{v_w}^{1} v^{-2}\, q(\theta, \widehat{\theta}_\pi, v)\, dv \tag{13}$$

where $q(\theta, \widehat{\theta}_\pi, v)$ denotes the quadratic risk $\mathbb{E}_{(\theta,v)}\| \widehat{\theta}_\pi - \theta\|^2$ based on the general model $\mathbf{M.2}(1,\eta,r)$ with $\sigma_p^2 = v$. Through the connecting equations the Kullback-Leibler risk can be viewed as an weighted aggregation of the square error risk.

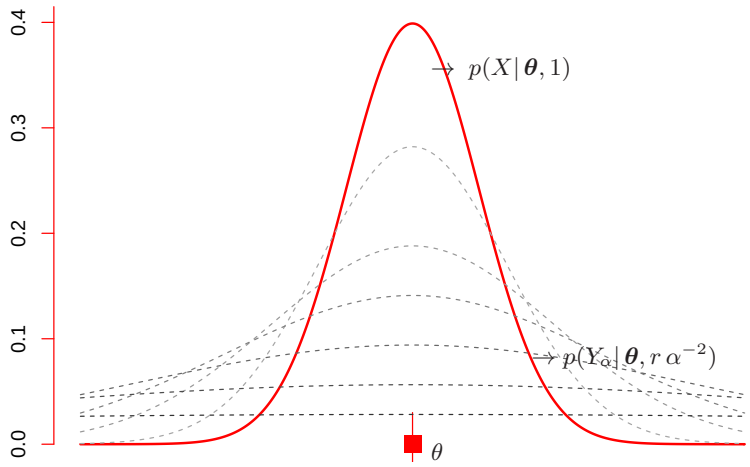

FIG 2. Path of Experiments:- *In red we have the true density of the observed random variable X around the unknown location θ. In different shades of gray (dark to light) we have respectively the density of Y_α for α equal 0.10, 0.20, 0.33, 0.50, 0.67 and 1.00. Hence the corresponding v equal 1.00, 0.98, 0.95, 0.89, 0.82 and 0.67. When, $\alpha = 1$, Y_α corresponds to the true future density around θ with known future variability $r = 2$.*

As the variance of Z varies from 1 to $v_w = (1+r^{-1})^{-1}$ it marks the gradual assimilation of the information in future Y to the existing information about θ in X through a *path of memoryless experiments* conducted separately for $\alpha \in [0,1]$. At each stage α along the path we observe $(X, Y[\alpha])$ where $Y[\alpha]$ is a Gaussian random variable around the true unknown location θ and with variability $r \cdot \alpha^{-2}$. Along the path, the information about the unknown location θ percolates through the sufficient statistics $\{Z[\alpha] : \alpha \in [0,1]\}$ where $Z[\alpha]$ is the UMVUE, of θ based on observing $(X, Y[\alpha])$. As α increases

in $[0,1]$, v decreases from 1 to v_w and we have,

$$Z[\alpha] = \frac{X[\alpha] + \alpha^2/r\, Y[\alpha]}{1 + \alpha^2/r} \quad \text{where } \alpha = r^{1/2}(v^{-1}-1)^{1/2} \in [0,1] \text{ and} \tag{14}$$

$$\widehat{\theta}^v_\pi = Z[\alpha] + v\,\nabla \log m_\pi(Z[\alpha], v) \tag{15}$$

By equation [27] (in the appendix) the plug-in risk $\rho^v_E(\theta, \widehat{\theta})$ equals $q(\theta, \widehat{\theta}_\pi, v)/(2v)$ and we see that $\rho(\theta, \widehat{p}_\pi)$ is equal to $\int_{v_w}^1 v^{-1}\rho^v_E(\theta, \widehat{\theta}_\pi)\, dv$ which implies that the predictive KL risk is a linearly weighted (according to precision) accumulation of the corresponding plug-in risk. Using the connecting equation [13] most of the calculations involving risks of different Bayes predictive densities would follow easily by using known evaluation of the quadratic risk of the corresponding posterior mean.

2.2. *Bayes-Minimax Method.*

We will explicitly solve for the equilibrium of the univariate minimax problem in $\mathbf{M.2}(1, \eta, r)$. Using the minimax theorem here, we see that over the class $\mathfrak{m}(\eta)$ the maximal univariate Bayes risk $\beta(\eta, r) = \sup\{B(\pi) : \pi \in \mathfrak{m}(\eta)\}$ is always less than the minimax risk $\rho_M(\eta, r)$. So, if we can produce:

1. a lower bound on $\beta(\eta, r)$ by considering the Bayes risk of a particular prior π_0 (say);
2. an upper bound on the minimax risk $\rho_M(\eta, r)$ by considering the $\max_\theta \rho(\theta, \widehat{p}_0)$ for a particular estimator $\widehat{p}_0$;
3. such that the lower bound and upper bound matches asymptotically as $\eta \to 0$;

we can conclude that $\beta(\pi_0)$ is the supremum Bayes risk as well as the minimax risk and π_0 is asymptotically least favorable and $\widehat{p}_0$ is a minimax strategy for the univariate predictive density estimation problem.

Once we have found the equilibrium for the univariate game, we extend the solution to the multivariate regime by following the general strategy outlined in Johnstone (2012, Section 4.11). Considering the class of exchangeable priors $\mathfrak{m}^e$ we can reduce the n–dimensional multivariate problem as repeated (n times) independent playing of the univariate minimax game and using the minimax theorem A.1 we can show that $R(n, s, r)$ would be less than $n\beta(s/n, r)$ (see Lemma 5.1). As detailed in Section 5, we can actually show that $R(n, s, r) \sim n\beta(s/n, r)$ would follow from concentration properties of the multivariate least favorable prior $\pi[n, s, r, 3]$. Following this scheme, we now calculate the asymptotic univariate minimax risk.

2.3. *The univariate asymptotic set-up.*

Here onwards we would further restrict our univariate parametric space to the non-negative orthant. The corresponding prior space would be $\mathfrak{m}^+(\eta) = \{\pi(\theta) : \pi(0) \geq 1 - \eta\}$. It would simplify exposition and the results easily generalizes over $\mathfrak{m}(\eta)$ by symmetrization.

To produce a lower bound on the maximal Bayes risk, we consider the class of all 2–point priors in $\mathfrak{m}^+(\eta)$. We will see that the 2–point version of the prior $\pi[\eta, r, 3]$ will attain the maximal Bayes risk where

$$\pi[\eta, r, 2](\theta) = \begin{cases} 0 & \text{with prob } 1 - \eta \\ \nu_\eta & \text{with prob } \eta \end{cases}$$

and ν_η is the positive root of the quadratic equation

$$\frac{1}{2}\, v_w^{-1}\, \nu^2 + v_w^{-1/2}\, \nu\, a = \log\left\{(1-\eta)\,\eta^{-1}\right\} \tag{16}$$

where $v_w = (1+r^{-1})^{-1}$ is the variance of the semi-futuristic random variable W and $a = (2\log\lambda_f)^{1/2}$ and $\lambda_f^2 = 2\, v_w\, \log\{(1-\eta)\,\eta^{-1}\}$.

To provide a detailed description of the univariate asymptotic regime we describe some fundamental quantities (functions of the sparsity level η) associated with our asymptotic calculations:

Universal Threshold: $\lambda_e = \left(2\log\{(1-\eta)\,\eta^{-1}\}\right)^{1/2}]$ is the universal threshold for point estimation of θ based on the past X only Donoho and Johnstone (1994a). The subscript 'e' emphasizes its estimative purpose. Later on, we will see that λ_e is also the optimal threshold in the predictive regime.

Ideal Predictive Threshold: $\lambda_f = (2v_w \log(1-\eta)\,\eta^{-1})^{1/2}$] is the universal threshold needed to devise minimax optimal threshold point estimate of θ based on observing the random variable W i.e both the past X and the complete future Y (which is equivalent to observing Y_α with $\alpha = 1$). The subscript f reflects its dependence on the future data.

Vantage Point: ν_η – the positive root of Equation [16] is the non-zero support point of the asymptotically least favorable 2–point prior in point estimation of θ under quadratic loss and noise variability v_w (compared with Johnstone (2012, Equation (8.36))) which again corresponds to location point estimation based on W. ν_η will be pivotal to our calculations. ν_η marks the beginning of the *Vulnerable Zone* which spans from $[\nu_\eta, \lambda_f]$. The calculation of the predictive risk on either side on ν_η displays the sparsity adjustments and uncertainty sharing dynamics.

Resolution Parameter, $a = (2\log\lambda_f)^{1/2}$*:* As $\eta \to 0$, we need to compute the asymptotic predictive risk as the true parameter θ moves along the non-negative axis. In the asymptotic regime, we can exactly quantify the risk

except at a few transition points (may be). However, the discontinuity of our analysis will only be limited to $\mathrm{O}(a)$–neighborhood around the transition points. In our calculations, a will generally arise an overshoot/undershoot parameter, e.g. Equation [16]. As in PE, a is of the order of $(\log\log\eta^{-1})^{1/2}$ and the risk can be accurately approximated in a resolution coarser than a. Now, note that Equation [16] reduces to $(v_w^{-1/2}\,\nu + a)^2 = \lambda_e^2 + a^2$ and so

$$\nu_\eta = \left(\lambda_f^2 + v_w\, a^2\right)^{1/2} - v_w^{1/2} a \geq \lambda_f - a v_w^{1/2}.$$

Thus $\nu_\eta \in (\lambda_f - a, \lambda_f)$. So as $\eta \to 0$, λ_f is quite close to the vantage point ν_η (in a–coarser resolution). Also, note that the ratio $\lambda_f : \lambda_e$ equals $v_w^{1/2} : 1$. Thus, as the future variability r decreases, the distance between λ_f and λ_e increases. Also as $\eta \to 0$, the threshold behaves as $\lambda_e \sim (2\log\eta^{-1})^{1/2}$. Attributes in the asymptotic regime are pictorially represented in Figure [3].

$\pi[\eta, r, 2]$ is a *sparse prior* in the sense that repeated sampling from the prior would lead us with a sparse signal as $\eta \to 0$. We will see that $\pi[\eta, r, 2]$ will be an asymptotically least favorable distribution for θ. To get a fair understanding of our strategies in the predictive regime we formulate the predictive density estimation problem as a two-person game between the Nature and a statistician.

2.4. *Predictive Two-Player Game and Equilibrium Strategies.*

In this predictive game, Nature chooses a probability distribution $\pi(\theta)$ from $\mathfrak{m}^+(\eta)$ for the location parameter θ. Then, a particular sample point θ_0 is generated from $\pi(\theta)$ and based on the signal θ_0 realizations X and Y contaminated with white noise would be produced: $X = \theta_0 + \epsilon_1$ and $Y = \theta_0 + r^{1/2}\epsilon_2$ where ϵ_1 and ϵ_2 are independent. The statistician sees only X and he knows about the sparsity restrictions and the data generation scheme. He has to come up with a density estimate for Y. It is to be noted that under sufficient concentration properties the complicated sparse high dimensional minimax prediction problem is equivalent to repeated playing of this simple 2–player game with fixed strategies (from both) over independent trials.

As $\eta \to 0$, a minimax strategy of this predictive game is given by the positive version $\widehat{p}[\eta, \mathrm{T}, \mathrm{CL}^+, \mathrm{U}]$ of $\widehat{p}[\eta, \mathrm{T}, \mathrm{CL}, \mathrm{U}]$ where

$$\widehat{p}[\eta, \mathrm{T}, \mathrm{CL}^+, \mathrm{U}](y|x) = \begin{cases} \widehat{p}\,(y|x; \pi[\eta, r, \mathrm{CL}^+]) & \text{if } X \leq \lambda_e \\ \widehat{p}\,(y|x; \pi_U) & \text{if } X > \lambda_e \end{cases}$$

and the $\pi[\eta, r, \mathrm{CL}^+]$ is a sparse discrete prior (cardinality of the support set equals $(K_\eta + 2)$ with K_η defined in Equation [17]) with $(1 - \eta)$ probability

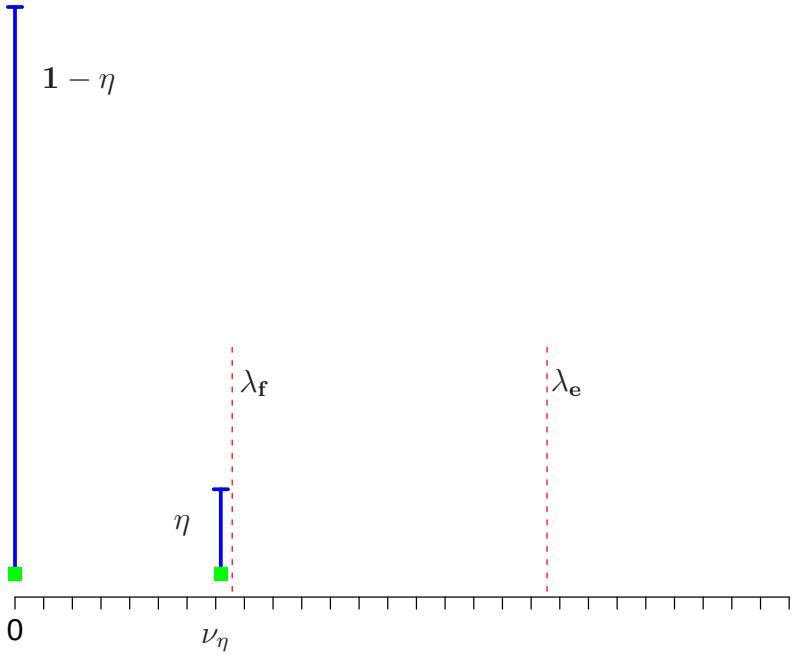

FIG 3. *The figure shows the support and probability allocation of the sparse* 2*–point prior* $\pi[\eta, r, 2]$ *along with the universal threshold* λ_e *and the ideal predictive threshold* λ_f. *The abscissa is graduated in a units and is drawn according to the scale with* $\eta = e^{-1000}$ *and* $r = 0.2$.

at 0 and sharing the remaining mass on a cluster of $(K_\eta + 1)$ support points. The non-zero support points approximately lies between λ_f and λ_e. The $(K_\eta + 1)$ non-zero support points start from $\mu_0 = \nu_\eta$ with ν_η given by the Equation [16] and span out in a geometric progression

$$\mu_i = (1+2r)^i \mu_0,\ i = 1, 2, \ldots, K_\eta \quad \text{and} \quad K_\eta = \max\left\{i : (1+2r)^i \mu_0 \leq \lambda_e + a\right\}.$$

As $\eta \to 0$, a first order approximation to the cardinality would be

$$K_\eta \sim \left\lfloor \frac{\log(\lambda_e/\lambda_f)}{\log(1+2r)} \right\rfloor = \left\lfloor \frac{\log(1+r^{-1})}{2\log(1+2r)} \right\rfloor. \tag{17}$$

The subscript in K_η would be dropped for simplicity. The non-zero support points are equally probable and in that way the cluster prior

$$\pi[\eta, r, \mathrm{CL}^+] = (1-\eta) \cdot \delta_0 + \frac{\eta}{K+1} \sum_{i=0}^{K} \delta_{\mu_i}(\theta)$$

is a probability distribution in $\mathfrak{m}^+(\eta)$ which is midway between the least favorable prior in PE based on X and $W = \text{UMVUE}(X, Y)$ respectively. The intermediation is marked by equal sharing of probability among the finite support points laid between λ_f and λ_e. A schematic representation of this mass allocation is presented in Figure [4]. The alignment of (spacing between) the support points is also intrinsic to the nature of the predictive problem and will be discussed later (in Section 4). Note that as $\eta \to 0$, $\pi[\eta, r, K]$ is a sparse prior and $\pi[\eta, r, \text{CL}^+] = \pi[\eta, r, 2]$ if $r > 0.42$.

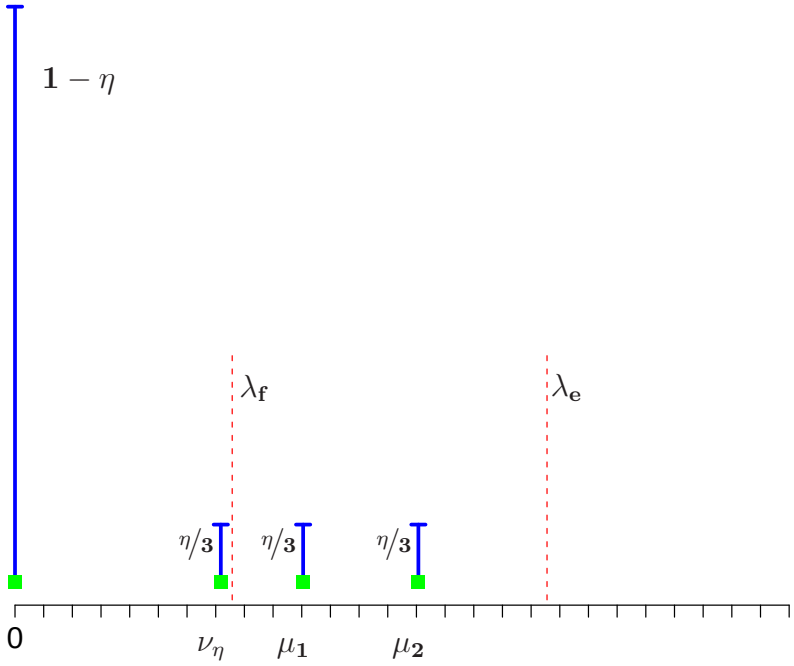

FIG 4. *The figure shows the support and probability allocation of the Cluster prior* $\pi[\eta, r, |Cl^+]$ *along with the universal threshold* λ_e *and the ideal predictive threshold* λ_f. *Here with* $r = 2$, *we have* 3 *equally likely non-zero support points at* $\mu_0 = \nu_\eta$, μ_1 *and* μ_2 *which constitute a geometric progression with common ratio* 1.4. *The abscissa is graduated in a units and is drawn to the scale of* $\eta = e^{-1000}$.

THEOREM 2.1. *As* $\eta \to 0$, $\pi[\eta, r, 2]$ *is an asymptotically least favorable prior for the univariate predictive game and* $\widehat{p}[\eta, T, CL+, U](y|x)$ *is an minimax estimator with the optimal asymptotic risk of* $(2r)^{-1}\,\eta\,\lambda_f^2$.

2.5. *Proof of Theorem 1.1.* As our set-up is symmetric it is enough to prove Theorem 2.1. We will use the the Bayes-Minimax strategy described before. So, we would calculate a lower bound on the Bayes risk of $\pi[\eta, r, 2]$ (see Lemma 2.1) and will produce a matching upper bound on the maximal risk of $\widehat{p}_T[\eta, \mathrm{T}, \mathrm{CL}^+, \mathrm{U}]$ (see Theorem 2.2). To interpret the result in terms of predictive game note that the statistician's choice for a density estimate of Y will involve information about θ_0 stored in X as well as the Gaussianity of the noise distribution. And, as the parametric form of the true future density is known an effective density estimate will depend on efficient estimation of θ_0 which in turn depends only on the sufficient statistics. In particular if X and Y were both observed an optimal point estimate of θ_0 based on the sufficient statistics W would produce an optimal density estimate (we will choose the plug-in version). So, for Nature, who aims to set the most difficult predictive set-up, the goal is essentially setting up the most difficult point estimation case for θ_0 based on the fact that X and Y are both observed. She apprehends that the statistician may produce a near accurate point prediction $\widehat{Y}$ of Y based on X. So, the worst possible prior distribution involve point estimation of θ_0 based on observing $(X, Y_\alpha)\big|_{\alpha=1}$. Let us denote a typical 2–point prior in $\mathfrak{m}^+(\eta)$ with its only non-zero support point at ν by

$$\pi_{2\mathrm{pt}}[\eta, \nu](\theta) = \begin{cases} 0 & \text{with prob } 1-\eta \\ \nu & \text{with prob } \eta \end{cases}$$

Note that the sparse two point prior $\pi[\eta, r, 2] = \pi_{2\mathrm{pt}}[\eta, \nu_\eta]$.

Lemma 2.1.

$$B(\pi[\eta, r, 2]) \geq \eta \frac{\lambda_f^2}{2r}\,(1 + o(1)) \quad \textit{as } \eta \to 0$$

Here, we provide an intuitive (and a bit non-rigorous) proof of the Lemma by using the connections with point estimation (PE) theory. We avoid the intricacies of overshoot term and present asymptotic arguments in the resolution higher than the $\mathrm{O}(a)$. In Section 3 we have rigorous technical proofs of the exact asymptotic behavior of the predictive risk of any 2–point priors in $\mathfrak{m}^+$. For those detailed calculation, the connecting equations can not help much as we still need to dig into the asymptotic subtleties of the PE regime. Lemma 2.1 and the other 2–point priors results will follow directly from those calculations.

Proof. To compute the predictive Bayes risk of $\pi[\eta, r, 2]$, we will use the Connecting equation [13] and known properties of the quadratic risk of

$\widehat{\theta}\big[\pi_{2\mathrm{pt}}[\eta,\nu]\big]$ – the posterior mean of the 2–point prior $\pi_{2\mathrm{pt}}[\eta,\nu]$.
From PE theory Johnstone (2012, Section 8.4) as $\eta \to 0$, the asymptotic quadratic risk $q\big(\theta, \widehat{\theta}\big[\pi_{2\mathrm{pt}}[\eta,\nu]\big], 1\big)$ of the Bayes estimate of $\pi_{2\mathrm{pt}}[\eta,\nu]$ in an estimative set-up with unit noise variance, has the following properties:

Property 1: Because of the very high mass at 0, the risk at 0 will be insignificant (lower than the order of $\eta \log \eta^{-1}$) and the dominant proportion of the Bayes risk will be from the non-zero support point ν.

Property 2: As $\eta \to 0$, the quadratic risk at the non-zero point ν will be of the order of ν^2 as long as $\nu^2 \leq \lambda_e^2 - c\,a\,\lambda_e$ with $c > 0$. Once ν exceeds λ_e the quadratic risk at ν becomes negligible compared to its peak value. Thus, the maximal first order asymptotic quadratic risk is attained when $\nu = \lambda_e$ and

$$q\big(\theta, \widehat{\theta}\big[\pi_{2\mathrm{pt}}[\eta,\nu]\big], 1\big) \sim \begin{cases} \nu^2 & \text{if } \nu^2 < \lambda_e^2 - c\,a\,\lambda_e \\ \mathrm{O}\,(1) & \text{if } \nu^2 \geq \lambda_e^2 + 2a\,\lambda_e \end{cases}.$$

Again, we know that the Gaussian estimative set up with noise variability v can be reduced to an unit variance problem by suitably scaling the observations as well as the location parameter by $v^{1/2}$. Posterior probabilities remain invariant to the transformation leading Bayes point estimates to be similarly scaled. And so, the quadratic risk of Bayes estimates under the variance stabilizing transformation is scaled by variability v, i.e.

$$q\big(\theta, \widehat{\theta}\big[\pi_{2\mathrm{pt}}[\eta,\nu]\big], v\big) = v \cdot q\big(v^{-1/2}\theta, \widehat{\theta}\big[\pi_{2\mathrm{pt}}[\eta, v^{-1/2}\nu]\big], 1\big).$$

Thus, while computing the predictive risk of the Bayes density estimate of $\pi[\eta, r, 2]$ at ν_η by the Equation [13] we have:

$$\begin{aligned} \rho\Big(\nu_\eta, \widehat{p}\big[\pi[\eta,r,2]\big]\Big) &= \int_{v_w}^1 \frac{1}{2v^2} q\big(\nu_\eta, \widehat{\theta}\big[\pi_{2\mathrm{pt}}[\eta,\nu_\eta]\big], v\big)\, dv \\ &= \int_{v_w}^1 \frac{1}{2v} q\big(v^{-1/2}\nu_\eta, \widehat{\theta}\big[\pi_{2\mathrm{pt}}[\eta, v^{-1/2}\nu_\eta]\big], 1\big)\, dv. \end{aligned}$$

Also, for all $v \in [v_w, 1]$ we have $(v^{-1/2}\nu_\eta)^2 \leq \lambda_e^2 - a\,\lambda_e$. So using the aforementioned Property 2, as $\eta \to 0$ we get

$$\rho\Big(\nu_\eta, \widehat{p}\big[\pi[\eta,r]\big]\Big) \geq \int_{v_w}^1 \frac{1}{2v} \times \frac{\nu_\eta^2}{v}\, dv = \nu_\eta^2 \int_{v_w}^1 \frac{1}{2v^2}\, dv = \frac{\nu_\eta^2}{2r} = \frac{\lambda_f^2}{2r}\,(1 + \mathrm{o}\,(1)) \text{ as } \eta \to 0$$

and the corresponding predictive Bayes risk satisfies

$$B(\pi[\eta,r,2]) \geq \eta \times r\Big(\nu_\eta, \widehat{p}\big[\pi[\eta,r]\big]\Big) \geq \frac{\lambda_f^2}{2r}\,(1 + \mathrm{o}\,(1)).$$

This completes the proof. □

Actually we can infer more about the prior $\pi[\eta, r]$ and like PE, here too we can show that the prior $\pi[\eta, r]$ is asymptotically least favorable among all 2–points priors.

LEMMA 2.2. *As $\eta \to 0$, $\pi[\eta, r, 2]$ maximizes the asymptotic Bayes risk in the class of all 2–point priors in $\mathfrak{m}^+$.*

PROOF. We know (by Property 1 described in Lemma 2.1) that the risk at the origin will have insignificant contribution (lower than the order of $\eta\lambda_e^2$) to the Bayes risk. Also, based on Property 2, for maximizing the risk at the non-zero support point the choice of ν is reduced to the set $\{\nu_k = k\lambda_e : k \in [0, 1]\}$. The predictive risk of $\pi_{2\text{pt}}[\eta, \nu_k]$ at the non-zero support point will be given by,

$$\begin{aligned} B\big(\pi_{2\text{pt}}[\eta, \nu_k]\big) &\sim \eta \cdot \rho\Big(\nu_\eta, \widehat{p}\big[\pi_{2\text{pt}}[\eta, \nu_k]\big]\Big) \\ &= \eta \int_{v_w}^1 \frac{1}{2v^2} q\big(\nu_\eta, \widehat{\theta}\big[\pi_{2\text{pt}}[\eta, \nu_k]\big], v\,\big)\, dv \\ &= \eta \int_{v_w}^1 \frac{1}{2v} q\big(v^{-1/2}\nu_\eta, \widehat{\theta}\big[\pi_{2\text{pt}}[\eta, v^{-1/2}\nu_k]\big], 1\,\big)\, dv. \end{aligned}$$

Now, as $\eta \to 0$, $q\big(v^{-1/2}\nu_\eta, \widehat{\theta}\big[\pi_{2\text{pt}}[\eta, v^{-1/2}\nu_k]\big], 1\,\big) \sim \begin{cases} v^{-1}\, k^2\, \lambda_e^2 & \text{if } v > k^2 \\ \mathrm{O}\,(1) & \text{if } v \le k^2 \end{cases}$

so, the asymptotic predictive Bayes risk is

$$\begin{aligned} B\big(\pi_{2\text{pt}}[\eta, \nu_k]\big) &= \eta\Big\{2^{-1}\, k^2\, \lambda_e^2 \int_{\max(k^2, v_w)}^1 v^{-2}\, dv\Big\}\big(1 + \mathrm{o}\,(1)\big) \\ &= k^2\left\{1 - (\max(k^2, v_w))^{-1}\right\}\eta\lambda_e^2/2\,\big(1 + \mathrm{o}\,(1)\big) \end{aligned}$$

which is maximized at $k = v_w^{1/2}$. Thus the Bayes risk is maximized for the 2–point prior whose non-zero support point is at λ_f. □

In this context, note that the optimal asymptotic predictive risk is always lower than the square error of point estimation of θ_0 based on X. This is because in the predictive set up (for Nature) the future Y could disclose additional information about θ_0. As such, we will see afterward that the ratio of the optimal predictive to estimated risk is v_w. For the two extreme cases as r approaches 0 and ∞ the ratio tends to 0 and 1 respectively. It validates our intuition about this predictive set-up. As with $r \to \infty$, even knowing Y will not provide any additional information about X. So, for

Nature the predictive problem will be as easy to set as the estimation one where only one sample X is explored. Similarly, as $r \to 0$, Y would disclose infinite amount of more information than X. Comparing the optimal risk in point estimation and the predictive regime we can say that predictive density estimation based on KL loss is an easier task than Point Estimation under quadratic loss. However, it does not say that prediction is easier than estimation.

On the contrary, constructing optimal predictive densities (for the statistician) is far more complicated than designing point estimates. The issues that need addressing are

Sparsity Regularization: The prior information of having at least $(1-\eta)$ mass at the origin has to incorporated.

Risk Diversification: The statistician does not see Y but can balance his ignorance about Y by sharing his future uncertainty.

As such, the interplay between sparsity-regularization needs and the dynamics of risk diversification (because of predictive purposes), requires to be explicitly tracked for calibrating optimal predictive schemes. We will see that the optimal strategies will lie outside the given parametric (Gaussian) family. However, there exists at least an asymptotic solution in the class of Gaussian mixture (finite) densities which is flexible enough to optimally balance the regularization vs diversification trade-off.

THEOREM 2.2. *As $\eta \to 0$, for any $r \in (0, \infty)$ we have*

$$\sup_{\pi \in \mathfrak{m}^+(\eta)} B(\pi, \widehat{p}[\eta, T, CL^+, U]) \leq \eta \lambda_f^2/(2r)(1 + o(1)).$$

The lemma is proved in section .4. For controlling the sparsity effect, the statistician can use a threshold density estimate. Threshold rules are a particular class of non-linear estimates which may be successfully employed to devise sparse minimax optimal estimates particularly in location estimation. The idea behind threshold rules is to use an unbiased estimate (generally unbiased or controlled bias) when the observed data is above the threshold and an adjusted one if the observation is below the threshold. Here we see that an optimal choice of threshold can depend entirely on the degree of sparsity η. As only X is observed the statistician is forced to use λ_e as the threshold. He can use the best invariant predictive density $\widehat{p}_U$ if the past observation X crosses λ_e. Below the threshold, his estimate has to account for both sparsity effect and risk sharing. The rationale for this choice rests on the fact that, because of the severe sparsity constraints a near zero density estimate is required to control the risk at the origin. If nature places the

entire remaining mass η between the parametric values $(0, \lambda_f]$, the statistician can perform under the optimal Bayes risk for this problem by using the zero density estimate $\phi(y|0, r)$. However, as the supremal Bayes risk is $\lambda_f^2/(2r)$ the zero estimator is unusable when the parametric value is greater than λ_f. Thresholding ensures that the predictive risk above λ_e is bounded. So, the need is to control the risk in $(\lambda_f, \lambda_e]$ by moving away from the zero estimator. But, this deformation should not be large to affect the risk at the origin and an approach would be to use the Bayes predictive density based on a prior with $(1 - \eta)$ probability at the origin (this leaves the sparsity restrictions untouched) and with η probability distributed in $(\lambda_f, \lambda_e]$.

2–Player Game and Equilibrium strategies

Nature: Will choose a distribution on θ which will make point estimation of θ, under quadratic loss and based on observing both the past X and future Y, the most difficult.

Statistician: Will use threshold estimators. He is forced to use λ_e as threshold as he only observes X. The idea is to use 0 estimator when $\theta < \lambda_f$ and share his risk for θ between λ_f and λ_e. A predictive strategy can be constructed by using a prior with probability $1 - \eta$ at 0 and share equally the remaining mass η among a finite chain of points covering λ_f and λ_e.

Decision Theoretic Evaluation Game

Through the above sharing policy we control the transition of the corresponding Bayes predictive density (and subsequently the threshold version) under $\theta \in (0, \lambda_e]$. The Bayes predictive density will be sufficiently close to $\phi(y|0, r)$ till $\theta < \lambda_f$ and thereafter it gradually shifts rightwards in way that the risk at any $\theta \in (\lambda_f, \lambda_e]$ is under the desired limit. The interval $(\lambda_f, \lambda_e]$ increases with decrease of r and the statistician is completely non-informative in that zone. As such $(\lambda_f, \lambda_e]$ can be regarded as his most **vulnerable zone**. A way to share his predictive risk across that zone would be to divide the probability η equally in a finite chain of point covering the interval. The non-informativeness in $(\lambda_f, \lambda_e]$ is reflected in uniform sharing of the future uncertainty. Quantizing the vulnerability in finite locations across the interval is pivotal. As soon as the parametric value θ crosses λ_f we need sharp transitions from the zero-estimator for which there is need for non-zero mass around λ_f. So, continuous sharing policies which are independent of the degree of sparsity η will not work.

In Section 6, we will see several efficient alignments of the chain. $\widehat{p}_T(.|x)$ is discontinuous at $x = \lambda_e$ and the number of Gaussian mixtures in $\widehat{p}_T$ and

their weights are based on the degree of sparsity η, the future volatility r and the past observation X. However, neither the Bayes density density corresponding to $\pi[\eta, r]$ attain minimax risk nor the cluster prior (which is the basis of $\widehat{p}_T$) is least favorable. But, based on the calculation we can trace an infinite support prior $\pi[\eta, r, \mathrm{INF}]$ on θ which attains supremal risk and produces minimax Bayes density estimate.

3. Maximal univariate Bayes risk of 2–point priors.

Here, we will work directly with the predictive loss. Calculation will involve deriving closed forms for the Bayes predictive densities. In the process we will show that properties similar to those stated in Lemma 2.1 for the quadratic loss also exist for predictive densities.

Posterior probabilities based on $\pi_{2\mathrm{pt}}[\eta, \nu]$ is given by

$$\pi_{2\mathrm{pt}}[\eta,\nu](\theta = 0|\, x) = \frac{(1-\eta)\,\phi(x)}{\eta\,\phi(x-\nu) + (1-\eta)\,\phi(x)} \quad \text{and}$$
$$\pi_{2\mathrm{pt}}[\eta,\nu](\theta = \nu|\, x) = 1 - \pi(\theta = 0|\, x).$$

And so, the corresponding Bayes predictive density is

$$\begin{aligned}\widehat{p}\big[\pi_{2\mathrm{pt}}[\eta,\nu]\big]\,(y|x) =& \frac{(1-\eta)\cdot\phi(x)\cdot\frac{1}{\sqrt{r}}\phi\big(\frac{y}{\sqrt{r}}\big) + \eta\cdot\phi(x-\nu)\cdot\frac{1}{\sqrt{r}}\phi\big(\frac{y-\nu}{\sqrt{r}}\big)}{(1-\eta)\cdot\phi(x) + \eta\cdot\phi(x-\nu)} \\ &= \frac{1}{\sqrt{r}}\cdot\phi\left(\frac{y}{\sqrt{r}}\right)\frac{(1-\eta) + \eta\cdot e^{\nu(x+y/r) - \frac{1+r}{2r}\cdot\nu^2}}{(1-\eta) + \eta e^{\nu x - \frac{1}{2}\nu^2}} \\ &= \phi(y|0,r)\times h_\nu(x,y).\end{aligned}$$

For doing calculations under strong sparsity, we found it most convenient to represent the Bayes predictive densities as tiltings of the zero-density. The tilt function $h_\nu(X, Y)$ is given by

$$h_\nu(x,y) = \big\{1+\eta\,(1-\eta)^{-1}e^{\nu x - \frac{1}{2}\nu^2}\big\}^{-1}\big\{1+\eta\,(1-\eta)^{-1}e^{\nu(x+y/r) - \frac{1+r}{2r}\cdot\nu^2}\big\} \tag{18}$$

with both the numerator and denominator being greater than unity which implies that their logarithms are always positive.

Now, from definition we have predictive risk at 0 as,

$$r\left(0, \widehat{p}\big[\pi_{2\mathrm{pt}}[\eta,\nu]\big]\right) = E_0\left(\log\left(\frac{\phi(Y|0,r)}{\widehat{p}\big[\pi_{2\mathrm{pt}}[\eta,\nu]\big]\,(Y|X)}\right)\right) = -E_0\left\{\log h_\nu(X,Y)\right\}$$

where the expectation is over both X and Y which are independent Gaussian with common mean (denoted in subscript) and known variances 1 and r

respectively. Also, note that though there is a negative sign the risk is always positive (as we have showed before that KL divergences are always positive).

The risk at ν is given by

$$\begin{aligned}
\rho\Big(\nu, \widehat{p}\big[\pi_{2\text{pt}}[\eta,\nu]\big]\Big) &= E_\nu\left(\log\left(\frac{\phi(Y|\nu,r)}{\widehat{p}\big[\pi_{2\text{pt}}[\eta,\nu]\big](Y|X)}\right)\right) \\
&= E_\nu\left(\log\left(\frac{\phi(Y|\nu,r)}{\phi(Y|0,r)}\right)\right) - E_\nu\{\log h_\nu(X,Y)\} \\
&= \frac{1}{2r} \times E_\nu(2\nu Y - \nu^2) - E_\nu\{\log h_\nu(X,Y)\} \\
&= \frac{\nu^2}{2r} - E_\nu\left\{\log h_\nu(X,Y)\right\}.
\end{aligned}$$

As $\eta \to 0$, for any 2–point prior with the very high probability $(1-\eta)$ at 0, the loss at the origin is always small irrespective of where the non-zero support ν is placed. We show in Lemma 3.1 that it remains bounded by η. At ν though, the asymptotic loss can be unbounded as $\eta \to 0$. We will chose an optimal ν maximizing this loss and the asymptotic maximal Bayes risk will be solely governed by risk at the non-zero support point in-spite of its low prior probability η.

Now as we move ν away from the origin, the Bayes density estimate at ν initially behaves like $\phi(.|0,r)$ giving rise to the first order asymptotic loss $\nu^2/2r$. In these cases the tilt function h_ν fails to sway the predictive density away from the origin when the true parametric value is ν. However, as we move ν further away from 0, $h_\nu(X,Y)$ will be successful in tilting the predictive density away from $\phi(.|0,r)$ and towards $\phi(y|\nu,r)$. Subsequently, the risk at ν will drop due to appreciable contribution from $E_\nu\left\{\log h_\nu(X,Y)\right\}$.

If the non-zero support point is ν_η (the positive root of the quadratic equation [16]) the tilt function is still inept enough not to cause any significant reduction in the first order asymptotic risk at ν_η. The proof follows directly from Lemma 3.2. So, with $\eta \to 0$, the first order asymptotic risk of the Sparse 2–point prior $\rho(\nu_\eta, \widehat{p}_{\pi[\eta,r,2]}) \geq \nu_\eta^2/(2r)\,(1+o(1))$ which in turn reproves Lemma 2.1 as

$$\begin{aligned}
B\big(\pi[\eta,r,2]\big) &= (1-\eta) \times r\big(0, \widehat{p}\big[\pi[\eta,r]\big]\big) + \eta \times r\big(\nu_\eta, \widehat{p}\big[\pi[\eta,r]\big]\big) \\
&\geq \eta\frac{\nu_\eta^2}{2r}\,(1+o(1)).
\end{aligned}$$

LEMMA 3.1. *For any $\eta \in [0,1)$ and $\nu \in [0,\infty)$ we have,*

$$r\left(0, \widehat{p}\big[\pi_{2pt}[\eta,\nu]\big]\right) \leq \eta(1-\eta)^{-1}.$$

PROOF. We use the representation of $h_\nu(X, Y)$ given by Equation [18] and so can assume that the logarithm of the numerator there is non-negative. Hence,

$$\begin{aligned}\rho\left(\nu, \widehat{p}\big[\pi_{2\mathrm{pt}}[\eta, \nu]\big]\right) &\leq E_0 \log\left\{1 + \eta\,(1-\eta)^{-1} e^{\nu X - \frac{1}{2}\nu^2}\right\} \\ &\leq \eta\,(1-\eta)^{-1} e^{-\nu^2/2}\, E_0\, e^{\nu X}\end{aligned}$$

by using the inequality $\log(1+x) \leq x$ which holds for all non-negative x. Again, as X is standard normal we have $E_0\left(\exp(\nu X)\right) = \exp(-\nu^2/2)$ and we have the required result. □

LEMMA 3.2. *For any $\eta \in [0,1]$ such that $a > 0$ and there exists a positive solution ν_η of Equation [16], we have*

$$\log(1-\eta) \leq E_\nu\left\{\log h_\nu(X, Y)\right\} \leq 1 + v_w^{-1/2} a^{-2} \text{ for all } \nu \in (0, \nu_\eta].$$

PROOF. We first show the upper bound. For that purpose we will use the representation of $h_\nu(X, Y)$ given by Equation [18] and so the logarithm of the denominator there is always positive and can be ignored while calculating the upper bound. We can rewrite the numerator in terms of the futuristic random variable $W = (1+r^{-1})^{-1}(X + Y/r)$ and its variance $v_w = (1+r^{-1})^{-1}$ as:

$$E_\nu \log\left\{1 + \eta\,(1-\eta)^{-1} \exp\left(v_w^{-1}\,\nu\,W - \frac{1}{2}\,v_w^{-1}\nu^2\right)\right\}$$

where $W \overset{d}{=} N(\nu, v_w)$. We change the measure to standard normal $Z = v_w^{-1/2}(W - \nu)$ and it results in

$$E_0 \log\left\{1 + \eta\,(1-\eta)^{-1} \exp\left(v_w^{-1/2}\,\nu\,Z + \frac{1}{2}\,v_w^{-1}\nu^2\right)\right\}$$

Now, as $\nu \leq \nu_\eta$ by Equation [16] we have,

$$\eta\,(1-\eta)^{-1} \exp\left(v_w^{-1/2}\,\nu\,Z + \frac{1}{2}\,v_w^{-1}\nu^2\right) = c_\nu\, \exp\left(v_w^{-1/2}\,\nu\,(Z-a)\right)$$

where c_ν is a constant and $c_\nu \in [0,1]$. Choosing $c_\nu = 1$ we get the upper bound,

$$E_\nu\left\{\log h_\nu(X, Y)\right\} \leq E_0 \log\left(1 + \exp\left(v_w^{-1/2}\,\nu\,(Z-a)\right)\right).$$

Now, we decompose the expectation of the random variable on R.H.S. conditioned on the event $\{Z > a\}$. When $Z \leq a$, the random variable (whose expectation is considered) is bounded by $\log 2$.
When $Z > a$, we use the naive bound: '$\log(1+x) \leq 1 + \log x$ if $x > 1$' for bounding the said random variable. Aggregating the two parts the ultimate bound would be

$$\begin{aligned} E_\nu \left\{\log h_\nu(X,Y)\right\} &\leq \log 2 \cdot P(Z \leq a) + \left[P(Z > a) + v_w^{-1/2}\, \nu E(Z-a)_+\right] \\ &\leq 1 + v_w^{-1/2}\, \nu E(Z-a)_+ \end{aligned}$$

and the truncated Gaussian expectation can be exactly computed as

$$E(Z-a)_+ = \int_a^\infty z\phi(z)\, dz - a\, \widetilde{\Phi}(a) = \phi(a) - a\, \widetilde{\Phi}(a) \leq a^{-2}\phi(a) \leq a^{-2}\lambda_f^{-1}$$

where the first inequality uses the result (28) about Mill's Ratio. As $\nu_\eta \leq \lambda_f$, for any $\nu \leq \nu_\eta$ we have $\nu E(Z-a)_+ \leq a^{-2}$. So, $E_\nu \left\{\log h_\nu(X,Y)\right\}$ is also bounded by $1 + v_w^{-1/2} a^{-2}$.

For lower bound we can similarly neglect the numerator of $h_\nu(X,Y)$ and so,

$$\begin{aligned} E_\nu \left\{\log h_\nu(X,Y)\right\} &\geq -\, E_\nu \log \left(1 + \eta\,(1-\eta)^{-1} e^{\nu X - \frac{1}{2}\nu^2}\right) \\ &\geq -\log \left(1 + \eta\,(1-\eta)^{-1} e^{-\frac{1}{2}\nu^2} E_\nu e^{\nu X}\right) \end{aligned}$$

which follows by Jensen's inequality. Noting, that X is standard Gaussian, the bound simplifies to $-\log\left(1+\eta\,(1-\eta)^{-1}\right) = \log(1-\eta)$. Hence, proved. □

By Lemma 3.1 and Lemma 3.2 the asymptotic behavior of $B\big(\pi_{2\text{pt}}[\eta,\nu]\big)$ can be characterized when ν varies in $[0,\nu_\eta]$. Next, we track $\rho\left(\nu, \widehat{p}\big[\pi_{2\text{pt}}[\eta,\nu]\big]\right)$ (and hence $B\big(\pi_{2\text{pt}}[\eta,\nu]\big)$ by using Lemma 3.1) when $\nu > \lambda_f$.
We will prove that the risk remains bounded as ν crosses the threshold λ_e (Lemma 3.3). In between λ_f and λ_e we can show that the risk is decreasing (Lemma 3.4). However, the descent is gradual and there is no abrupt transition in the first order risk before λ_e. Thus, the maximal Bayes risk for 2–point prior is attained around $\nu = \lambda_f$ and we have effectively characterized the first order behavior of the risk with ν varying along the positive axis in resolution of a units. As such, as $\eta \to 0$

$$\rho\left(\nu, \widehat{p}\big[\pi_{2\text{pt}}[\eta,\nu]\big]\right) \sim \begin{cases} \nu^2/2r & \text{if } \nu \leq \nu_\eta \\ \text{decreasing} & \text{if } \lambda_f + (2\, v_w^{-1})^{1/2} a \leq \nu \leq \lambda_e - 2a \\ \text{O}\,(1) & \text{if } \nu > \lambda_e + (2 v_w^{-1})^{1/2} a \end{cases} .$$

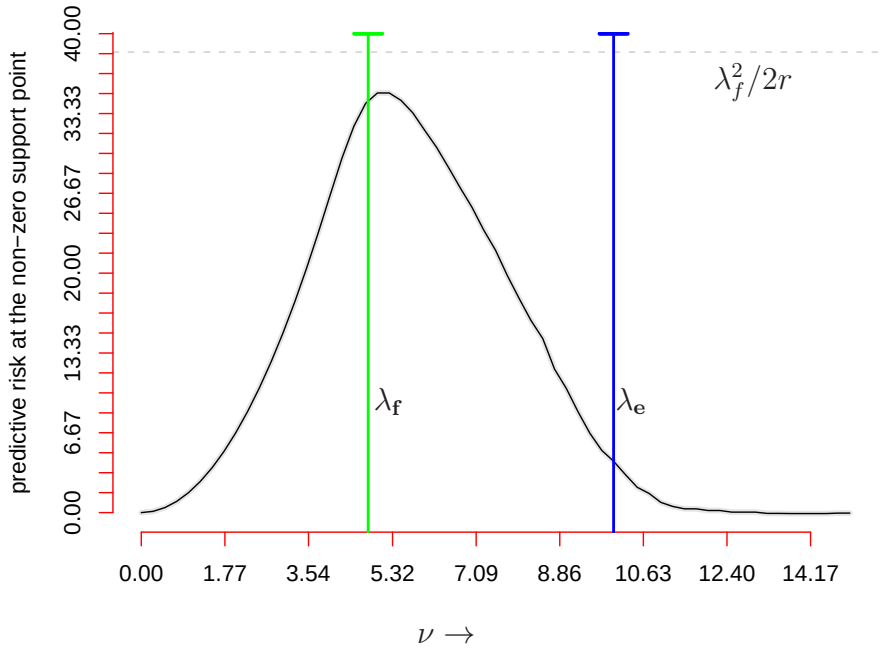

FIG 5. *The plot shows the risk* $\rho\left(\nu, \widehat{p}\left[\pi_{2pt}[\eta, \nu]\right]\right)$ *of the Bayes predictive density from the two point prior* $\pi_{2pt}[\eta, \nu]$ *at the non-zero support point* ν *as* ν *is moved along the positive axis.* λ_f *and* λ_e *are marked by vertical lines and the horizontal gray line represents the first order maximal risk of* $\lambda_f^2/2r$*. Reflecting the resolution of our asymptotic calculations the abscissa is ticked at multiples of* a *while the ordinate is marked in multiple of* $1/(2r)$ *units to represent change in order of quadratic loss. The figure is actually drawn according to scale with* $\eta = e^{-50}$*,* $r = 0.3$ *producing* $\lambda_f = 4.8$*,* $\lambda_e = 10$*,* $a = 1.77$*.*

We do not quantify the rate of descent of $\rho\left(\nu, \widehat{p}\big[\pi_{2\mathrm{pt}}[\eta,\nu]\big]\right)$ though it can be approximated from our proof of Lemma 3.4. In Figure [5], we trace the asymptotic risk by Monte Carlo simulation. It depicts the gradual descent which compared to PE regime is a contrast.

LEMMA 3.3. *For any $\eta \in (0,1)$ such that $a > 0$ we have,*

$$\rho\left(\nu, \widehat{p}\big[\pi_{2pt}[\eta,\nu]\big]\right) \leq (2\,a\,r)^{-1} + \eta(1-\eta)^{-1}\log 2 \text{ for all } \nu > \lambda_e + (2v_w^{-1})^{1/2}a.$$

PROOF. We know that $\rho\left(\nu, \widehat{p}\big[\pi_{2\mathrm{pt}}[\eta,\nu]\big]\right) = \nu^2/2r - E_\nu\{\log h_\nu(X,Y)\}$. As before we can standardize the measure to standard Gaussian. Because of the logarithm the numerator and denominator of h_ν separates and can be standardized simultaneously. After standardization, we have:

(19)
$$\rho\left(\nu, \widehat{p}\big[\pi_{2\mathrm{pt}}[\eta,\nu]\big]\right) = \frac{\nu^2}{2r} - E_0 \log\left\{\frac{1+\eta\,(1-\eta)^{-1}\exp\left(v_w^{-1/2}\,\nu\,Z + \frac{1}{2}\,v_w^{-1}\nu^2\right)}{1+\eta\,(1-\eta)^{-1}\exp\left(\nu\,Z + \frac{1}{2}\nu^2\right)}\right\}$$

(20)
$$= \frac{\nu^2}{2r} - E_0 \log\left\{\frac{1+\exp\left(v_w^{-1/2}\,\nu\,Z + \frac{1}{2}\,v_w^{-1}\nu^2 - \frac{1}{2}\lambda_e^2\right)}{1+\exp\left(\nu\,Z + \frac{1}{2}\nu^2 - \frac{1}{2}\lambda_e^2\right)}\right\}$$

by substituting $\eta(1-\eta)^{-1}$ by $\exp(-\lambda_e^2/2)$. Note that $\nu > \lambda_e + (2v_w^{-1})^{1/2}a$ implies $2\Delta = \nu^2 - 2\nu b - \lambda_e^2 > 0$ where $b = (4\,v_w^{-1}\log\nu)^{1/2}$ and hence the expectation above can be rewritten as

$$E_0 A(Z) \text{ where } A(Z) = \log\left\{\frac{1+\exp\left(v_w^{-1/2}\,\nu\,(Z+bv_w^{1/2})\right)\cdot\exp(\Delta)\cdot\exp(\nu^2/2r)}{1+\exp\left(\nu\,(Z+b)\right)\cdot\exp(\Delta)}\right\}$$

where Z is standard normal distribution. Again, when $Z > -b\,v_w^{1/2}$, $A(Z) \geq \nu^2/2r$. And $A(Z)$ is negative iff $Z < -\nu$ which leads us to:

$$E_0 A(Z) \geq \frac{\nu^2}{2r}\cdot P\big(Z > -b\,v_w^{1/2}\big) - E_0\left(\left\{1+\exp\left(\nu\,Z + \frac{1}{2}\nu^2 - \frac{1}{2}\lambda_e^2\right)\right\}\mathbb{I}[Z<-\nu]\right)$$
$$\geq \frac{\nu^2}{2r}\cdot\Phi\big((4\log\nu)^{1/2}\big) - \log 2\;\cdot\;\Phi(-\nu)$$

as the random variable within the expectation is always less than $\log 2$. Eventually, we arrive at

$$\rho\left(\nu, \widehat{p}\big[\pi_{2\mathrm{pt}}[\eta,\nu]\big]\right) \leq \frac{\nu^2}{2r}\cdot\widetilde{\Phi}\big((4\log\nu)^{1/2}\big) + \log 2\cdot\widetilde{\Phi}(\nu) \leq (2\,a\,r)^{-1} + \eta(1-\eta)^{-1}\log 2$$

where the second inequality follows from Equation [28]. Thus, we get the result. □

LEMMA 3.4. *As* $\eta \to 0$ *and* $\nu^2 \in \left(\left\{ \lambda_f + (2\, v_w^{-1})^{1/2} a \right\}^2,\ \lambda_e^2 - 2\, a\, \lambda_e \right)$ *the risk* $\rho\left(\nu, \widehat{p}\left[\pi_{2pt}[\eta, \nu]\right]\right)$ *is dominated by a decreasing function of* ν *which is bounded above by* $\lambda_f^2/(2r)$.

PROOF. The risk is given by Equation [19]. Similarly as before we can show that as long as $\nu^2 < \lambda_e^2 - 2a\lambda_e$ the contribution from the denominator is insignificant

$$E_0 \log \left\{ 1 + \exp\left(\nu\, Z + \frac{1}{2}\nu^2 - \frac{1}{2}\lambda_e^2\right) \right\} \leq O(1)$$

and whenever $\nu \geq \lambda_f + (2\, v_w^{-1})^{1/2} a$ we have,

$$E_0 \log \left\{ 1 + \exp\left(v_w^{-1/2}\, \nu\, Z + \frac{1}{2}\, v_w^{-1}\nu^2 - \frac{1}{2}\lambda_e^2\right) \right\} \geq \frac{1}{2}\left(\lambda_e^2 - v_w^{-1}\nu^2\right) \cdot \Phi\left(\sqrt{2}a\right).$$

So, ultimately we will have

$$\rho\left(\nu, \widehat{p}\left[\pi_{2\mathrm{pt}}[\eta, \nu]\right]\right) \leq \frac{1}{2}\left(\lambda_e^2 - \nu^2\right) + \frac{\nu^2}{2r}\, \widetilde{\Phi}\left(\sqrt{2}a\right) \leq \frac{1}{2}\left(\lambda_e^2 - \nu^2\right) + 1$$

which is decreasing in ν and bounded above by $\lambda_f^2/2r$. □

It will be seen that to prove Theorem we only need Lemma 2.1 for the lower bound. In that regard, the extensive calculations in this section may seem to be an expensive regression from the objective. However, these results not only provide more intuition about the predictive regime but will also help to follow the risk calculations in the next section where the risk of the density estimate from a multi (K) point prior compounded with thresholding complications is evaluated.

4. Minimax upper bound.

To simplify notations the univariate strategy $\widehat{p}[\eta, \mathrm{T}, \mathrm{CL}^+, \mathrm{U}]$ and the cluster prior $\widehat{p}[\eta, r, \mathrm{CL}^+]$ will be abbreviated as $\widehat{p}_T$ and $\pi[\eta, r, K]$ through out this section. K is the number of support point in the cluster prior and is a function of η and r. As mentioned before, the $\widehat{p}_T$ is governed by 2 major effects: thresholding and risk diversification. The risk diversification procedure involves (a) probability allocation (b) alignment of non-zero support points across the vulnerable zone, which in our $\widehat{p}_T$ is characterized by $\pi[\eta, r, K]$. In this section, the risk calculations of $\widehat{p}_T$ are carried out in a manner such that they can be easily generalized for other reasonable discrete probability sharing schemes (with $1 - \eta$ probability at the origin). In particular, we incorporate the peculiar alignment of the non-zero support points of $\pi[\eta, r, K]$ only toward the end of the section and the results before Lemma 4.4 will be reused in Section 6 to display other feasible sharing schemes.

Proof of Theorem 2.2. We characterize the Bayes predictive density for the Cluster prior. Using Bayes formula, the posterior distribution of $\pi[\eta, r, K]$ is given by:

$$\pi[\eta,r,K](0\,|x) = \frac{(1-\eta)\;\phi\,(x)}{(1-\eta)\;\phi\,(x) + \eta/(K+1)\sum_{i=0}^{K}\phi\,(x-\mu_i)}$$
$$\pi[\eta,r,K](\mu_j\,|x) = \frac{\eta/(K+1)\;\phi\,(x-\mu_j)}{(1-\eta)\;\phi\,(x) + \eta/(K+1)\;\sum_{i=0}^{K}\phi\,(x-\mu_i)}, \quad j=0,\ldots,K$$

and the predictive density $\widehat{p}\big[\,\pi[\eta,r,K]\,\big]\big(y\big|x\big)$ is given by

$$\pi[\eta,r,K](0\,|x)\;\cdot\;\phi(y|0,r) + \sum_{i=0}^{K}\pi[\eta,r,K](0\,|x)\;\cdot\;\phi(y|\mu_i,r)$$
$$= \frac{(1-\eta)\,\phi(x)\,\phi(y|0,r) + \eta/(K+1)\sum_{i=0}^{K}\phi(x-\mu_i)\,\phi(y|\mu_i,r)}{(1-\eta)\,\phi(x) + \eta/(K+1)\sum_{i=0}^{K}\phi(x-\mu_i)}$$

which like the 2-point prior case can be rewritten as a tilt function acting on the zero density $\widehat{p}\big[\pi[\eta,r,K]\big] = \phi(y|0,r)\times h\big[\,\pi[\eta,r,K]\,\big](x,y)$ where

(21)

$$h\big[\,\pi[\eta,r,K]\,\big](x,y) = \frac{1+\frac{\eta(1-\eta)^{-1}}{K+1}\sum_{i=0}^{K}\exp\left\{\mu_i(x+\frac{y}{r}) - \frac{1}{2}(1+r^{-1})\mu_i^2\right\}}{1+\frac{\eta(1-\eta)^{-1}}{K+1}\sum_{i=0}^{K}\exp\left\{\mu_i x - \frac{1}{2}\mu_i^2\right\}}\,.$$

Observing $X=x$, its predictive loss at the point θ is expressed as :

$$L\Big(\theta,\widehat{p}\,\big[\,\pi[\eta,r,K]\,\big]\,\big(.\,|X=x\big)\Big) = \mathbb{E}_\theta\left[\log\left(\frac{\phi(Y|\theta,r)}{\widehat{p}[\,\pi[\eta,r,K]\,](Y|x)}\right)\right]$$
$$= \frac{\theta^2}{2r} - \mathbb{E}_\theta\Big(h\big[\,\pi[\eta,r,K]\,\big]\,(x,Y)\Big)$$

where the discounted location distance $\theta^2/2r$ appears due to the predictive loss between $\phi(Y\,|\,\theta,r)$ and $\phi(Y\,|\,0,r)$.
We would need to study the behavior of the conditional expectation as well as the unconditional $\mathbb{E}_\theta\Big(\,h\big[\,\pi[\eta,r,K]\,\big]\,(\,X,\,Y\,)\,\Big)$ (the expectation is over both X and Y) in details. For that purpose it will be helpful to change the measure to central Gaussian with X and Y having variances 1 and r respectively

$$\mathbb{E}_\theta\Big(h\big[\,\pi[\eta,r,K]\,\big]\,(X,Y)\Big) = \mathbb{E}_0\big(N_\theta(X,Y) - D_\theta(X)\big)$$

where N_θ and D_θ are the logarithms of the numerator and denominator of $h\big[\,\pi[\eta, r, K]\,\big]$ and

$$N_\theta(x,y) = \log\left[1 + \frac{\eta(1-\eta)^{-1}}{K+1}\sum_{i=0}^{K}\exp\left\{\mu_i\left(x+\frac{y}{r}\right) - \frac{1}{2}v_w^{-1}\mu_i^2 + v_w^{-1}\mu_i\theta\right\}\right]$$
$$= \log\left[1 + \frac{\eta(1-\eta)^{-1}}{K+1}\sum_{i=0}^{K}\exp\left\{v_w^{-1}\left(\mu_i w - \frac{1}{2}\mu_i^2 + \mu_i\theta\right)\right\}\right].$$

where W is the semi-futuristic random variable and $W \sim N(0,1)$. Now, using the fact that $\eta(1-\eta)^{-1}$ actually equals $\exp(v_w^{-1}\lambda_f^2/2)$ (by Equation [16])we have,

(22a)
$$N_\theta(x,y) = \log\left[1 + \frac{1}{K+1}\sum_{i=0}^{K}\exp\left\{v_w^{-1}\left(\mu_i w - \frac{1}{2}\mu_i^2 + \mu_i\theta - \frac{1}{2}\lambda_f^2\right)\right\}\right]$$

(22b)
$$= \log\left[1 + (K+1)^{-1}\sum_{i=0}^{K}\exp\left\{v_w^{-1}\Gamma_i(\theta)\right\} \times \exp\left\{v_w^{-1}\mu_i(W+a)\right\}\right]$$

(22c)
$$\text{where } \Gamma_i(\theta) = \mu_i\theta - \frac{1}{2}\mu_i^2 - a\mu_i - \frac{1}{2}\lambda_f^2$$

Thus when $v_w^{-1}\mu_i(W+a) \geq 0$ for all $i \in \{0,1,\ldots,K\}$ a very naive lower bound is

$$N_\theta(x,y) \geq v_w^{-1}\Gamma(\theta) - \log(K+1) \text{ where } \Gamma(\theta) = \max_{i=0}^{K}\Gamma_i(\theta). \tag{22d}$$

Similarly, for the denominator we have

$$D_\theta(x) = \log\left[1 + \frac{\eta(1-\eta)^{-1}}{K+1}\sum_{i=0}^{K}\exp\left\{\mu_i x - \frac{1}{2}\mu_i^2 + \mu_i\theta\right\}\right]. \tag{22e}$$

Next we calculate the risk of our threshold estimate

$$\widehat{p}_T(y|x) = \begin{cases} \widehat{p}\big(y|x;\pi[\eta,r,K]\big) & \text{if } X \leq \lambda_e \\ \widehat{p}\big(y|x;\pi_U\big) & \text{if } X > \lambda_e \end{cases}$$

We will later see that the threshold estimator $\widehat{p}_T(.|x)$ is discontinuous at $x = \lambda_e$. However, modifications of $\widehat{p}_U$ can be used to incorporate continuity correction. We are interested in finding the maximum risk of $\widehat{p}_T$. However,

due to high prior probability concentration at 0 we have to treat the risk at the origin separately. Depending on X the loss of the threshold estimate is given by:

$$L\Big(\theta, \widehat{p}\big[\pi[\eta, r, K]\big](.|x)\Big) = \frac{\theta^2}{2r} - \mathbb{E}_0 N_\theta(x-\theta, Y) + D_\theta(x-\theta) \qquad \text{if } x \le \lambda_e$$

$$L\Big(\theta, \widehat{p}\big[\pi_U\big](.|x)\Big) = \frac{1}{2}\big(\log(1+r^{-1}) - (1+r)^{-1}\big) + \frac{(x-\theta)^2}{2(1+r)} \qquad \text{if } x > \lambda_e$$

where the loss of $\widehat{p}_{\pi_U}$ follows from the risk calculations of linear estimators (see Appendix). Now, averaging over the observed data X, the risk of $\widehat{p}_T$ will be given by:

$$\rho(\theta, \widehat{p}_T) = \rho_B(\theta) + \rho_A(\theta)$$

where $\rho_A(\theta)$ is the risk when X crosses the threshold

$$\rho_A(\theta) = \frac{1}{2}\big(\log(1+r^{-1}) - (1+r)^{-1}\big)\, P_\theta(X \ge \lambda_e) + \frac{\mathbb{E}_\theta[(X-\theta)^2 I_{\{X \ge \lambda_e\}}]}{2(1+r)}$$

and the component of risk from below the threshold is

$$\begin{aligned}
\rho_B(\theta) &= \frac{1}{2r}\big[\,\rho_{B,1}(\theta) - \rho_{B,2}(\theta) + \rho_{B,3}(\theta)\,\big] \quad \text{where,}\\
\rho_{B,1}(\theta) &= \theta^2\,\Phi(\lambda_e - \theta)\\
\rho_{B,2}(\theta) &= 2r\mathbb{E}_0\left[N_\theta(X,Y) I_{\{X \le \lambda_e - \theta\}}\right]\\
\rho_{B,3}(\theta) &= 2r\mathbb{E}_0\left[D_{_\theta}(X) I_{\{X \le \lambda_e - \theta\}}\right].
\end{aligned}$$

As we have mentioned before, due to the very high probability of the parameter to concentrate at the origin we need to bound both $\rho_A(0)$ and $\rho_B(0)$ with high precision. Note that, by definition $N_\theta(X,Y) \ge 0$ and so $\rho_D(0) \le \mathbb{E}_0 D_0(X)\, I_{\{X \le \lambda_e\}}$ which again equals $\eta + o(\eta)$ as $\eta \to 0$ by Lemma 4.1. Though $\rho_B(0)$ will be significantly larger than $\rho_A(0)$, it will not be enough to carry the risk at 0 above the maximum value as,

$$\rho_A(0) \le \frac{1}{2}\log\left(1+r^{-1}\right)\tilde{\Phi}(\lambda_e) + \frac{1}{2(1+r)}\mathbb{E}_0\big(X^2 I_{\{X \ge \lambda_e\}}\big) \tag{23}$$

$$\le \frac{1}{2}\left(\log\frac{r+1}{r}\cdot\tilde{\Phi}(\lambda_e) + \frac{1}{1+r}\lambda\phi(\lambda_e)\right) \tag{24}$$

$$= O(\eta\lambda) \quad [\text{as } \phi(\lambda_e) = \eta/\sqrt{2\pi}] \tag{25}$$

And, from calculation involving the risk of hard threshold estimators we know that $\mathbb{E}_0\big(X^2 I_{\{X \ge \lambda_e\}}\big) = \lambda_e\phi(\lambda_e)$ Johnstone (2012, Equation (8.15))

and the using the result in equation [28] we have $\rho_B(0) = O(\eta\lambda)$. Hence the risk at 0 stays well below the maximal value.

Next we need to produce an upper bound on the maximum risk at any non-zero parametric point. While working with $r(0, \widehat{p}_T)$ we saw that the significant contribution came from $r_B(0)$. Outside the origin, the maximal predictive risk is governed solely by $\rho_B(\theta)$ which can be unbounded as $\eta \to 0$ while $\rho_A(\theta)$ remains bounded by $2^{-1}\log(1 + r^{-1}) + (1 + r)^{-1}$.
Now we trace the behavior of $\rho_B(\theta)$ as θ varies. It will vividly demonstrate how the dynamics of sharing future risk can be co-ordinated with sparsity prior information.

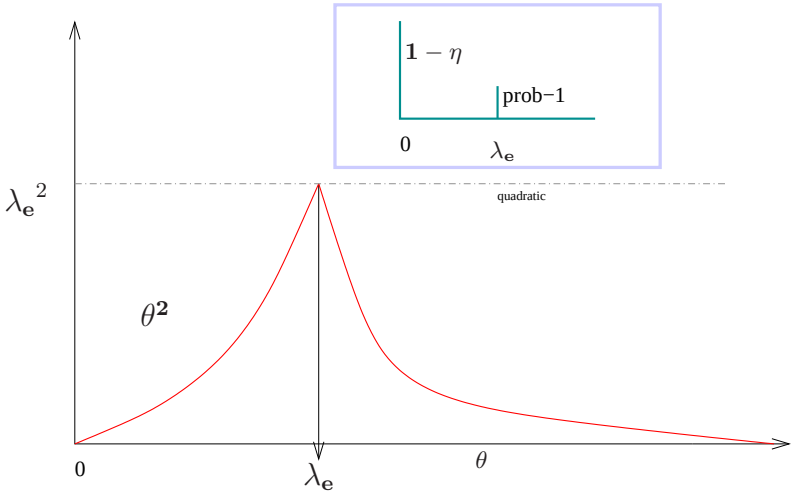

FIG 6. *Schematic Diagram of the Quadratic Risk in Minimax Sparse Point Estimation.*

$\rho_{B,1}(\theta)$ is the dominant portion of quadratic risk of the Hard threshold point estimator of θ. From point estimation theory we know that it behaves as θ^2 until the threshold λ_e and then shrinks to 0 with a steep decent (see Figure [6]). $\rho_{B,2} - \rho_{B,3}$ is the diversification or aggregation effect. $\rho_{B,3}$ being based on X entirely will be insignificant before λ_e due to sparsity and negligible thereafter due to thresholding effect. It is technically proved in Lemma 4.1. So, $\rho_{B,2}$ portrays the diversification effect. It is dormant before λ_f. In between $\lambda_f + a$ and λ_e, ρ_B^2 produces significant contribution and is effective in bringing the predictive risk $r(\theta, \widehat{p}_T)$ below $\lambda_f^2/2r$. The technical details of the following first order behavior of ρ_B is carried out in a serially

in the Lemma 4.2, Lemma 4.3 and Lemma 4.4:

$$\rho_{B,1}(\theta) \sim \begin{cases} \theta^2 & \text{if } \theta < \lambda \\ 0 & \text{if } \theta \geq \lambda + a \end{cases}$$
$$\rho_{B,2}(\theta) \sim \begin{cases} 2g(\theta) & \text{if } \theta \in [\mu_0 + a, \lambda + a] \\ 0 & \text{otherwise} \end{cases}$$
$$\rho_{B,3}(\theta) \sim 0 \text{ for all } \theta$$
$$\text{And, } \rho_{B,1}(\theta) - 2g(\theta) \leq \mu_0^2 \text{ for } \theta \in [\mu_0 + a, \lambda + a].$$

So we have,

$$\sup_{\theta > 0} r(\theta, \widehat{p}_T) \leq \frac{\lambda_f^2}{2r} + o(\mu_0^2) \quad \text{where } a = \sqrt{2 \log \mu_0}$$

and the minimax predictive risk of $\widehat{p}_T$ is

$$\sup_{\pi \in \mathfrak{m}^+(\eta)} \int r(\theta, \widehat{p}_T)\, \pi(\theta)\, d\theta \leq (1 - \eta)\, r(\theta, \widehat{p}_T) + \eta \sup_{\theta \geq 0} \rho_T(\mu, \lambda) \leq \eta \frac{\mu_0^2}{2r}(1 + o(1)).$$

Figure [6] and Figure [1] show schematic diagrams of the univariate risk plot of threshold rules in the PE and density estimation framework. The sole difference between the two regimes is the reduction $\rho_{B,2}(\theta)$ in the predictive risk in the vulnerable zone $[\lambda_f, \lambda_e]$. Plug-in density estimates fail to attain this risk reduction and have risk properties like optimal threshold estimates in PE. To obtain the risk reduction we need to have diversified predictive schemes. Lemma 4.3 provides a crude lower bound on the decrement and Lemma 4.4 shows that it is sufficient enough to attain first order optimality. Figure [7] contains Monte-Carlo simulation of the different predictive risk curves.

LEMMA 4.1. *For any $\eta \in (0, 1)$ such that λ_e is well defined and greater than 1, we have*

(a) $\mathbb{E}_0 \left\{ D_0(X)\, \mathbb{I}[X \leq \lambda_e] \right\} \leq -\log(1 - \eta)$
(b) $\mathbb{E}_0 \left\{ D_\theta(X)\, \mathbb{I}[X \leq \lambda_e - \theta] \right\} \leq \log 2$

PROOF. The first inequality follows directly from Jensen's inequality,

$$\begin{aligned}\mathbb{E}_0\left\{D_0(X)\,\mathbb{I}[X\le\lambda_e]\right\} &\le \log\left[\mathbb{E}_0\left\{\left(1+\frac{\eta\,(1-\eta)^{-1}}{K+1}\sum_{i=0}^{K}e^{\mu_i X-\frac{1}{2}\mu_i^2}\right)\mathbb{I}[X\le\lambda_e]\right\}\right]\\ &\le \log\left[1+\frac{\eta\,(1-\eta)^{-1}}{K+1}\sum_{i=0}^{K}\mathbb{E}_0\left(e^{\mu_i X-\frac{1}{2}\mu_i^2}\right)\right]\\ &= \log\left(1+\eta\,(1-\eta)^{-1}\right) = -\log(1-\eta)\end{aligned}$$

For the second inequality, as $D_\theta(X)$ is an increasing function of X we have

$$\mathbb{E}_0\left\{D_\theta(X)\,\mathbb{I}[X\le\lambda_e-\theta]\right\} \le D_\theta(\lambda_e-\theta)\,\Phi(\lambda-\theta) \quad \text{where,}$$
$$D_\theta(\lambda-\theta) = \log\left(1+\frac{\eta(1-\eta)^{-1}}{K+1}\sum_{i=0}^{K}e^{\mu_i\lambda_e-\frac{1}{2}\mu_i^2}\right).$$

Also, as for each $i\in\{0,1,\ldots,K\}$ we have $0\le\mu_i\le\lambda_e+a$, and so the maximum value of $\mu_i\lambda_e-\mu_i^2/2$ is at most $\lambda^2/2$ for all $i\in\{0,1,\ldots,K\}$ which would imply that $D_\theta(\lambda-\theta)\le\log\left(1+\eta(1-\eta)^{-1}\exp(\lambda_e^2/2)\right)\le\log 2$. Hence, proved. □

LEMMA 4.2. *For any $\theta\ge\lambda_e+a$, $\rho_{B,1}(\theta)\le O(\lambda_f)$ as $\eta\to 0$.*

PROOF. This Lemma follows from the risk calculations of threshold point estimates. Taking derivative, we see

$$\rho'_{B,1}(\theta) = 2\theta\,\Phi(\lambda_e-\theta)-\theta^2\phi(\lambda_e-\theta) \text{ and so for } t\ge 0$$
$$\rho'_{B,1}(\lambda_e+t) = 2(\lambda_e+t)\Phi(t)-(\lambda_e+t)^2\phi(t)\le\left(2t^{-1}-(\lambda_e+t)\right)(\lambda_e+t)\,\phi(t)$$

where the inequality follows from Equation [28]. Hence, for all $t\ge 2\,\lambda_e^{-1}$, $\rho'_{B,1}(\lambda_e+t)$ is negative. As, $\eta\to 0$, we have $a\ge 2\,\lambda_e^{-1}$ which implies that $\rho_1(\theta)$ is a decreasing function of θ as $\theta>\lambda_e+a$. So, as $\eta\to 0$,

$$\sup_{\theta\ge\lambda_e+a}\rho_{B,1}(\theta)\le(\lambda_e+a)^2\tilde{\Phi}(a)\le(\lambda+a)^2\,a^{-1}\phi(a)=(\lambda+a)^2a^{-1}\lambda_e^{-1}=O(\lambda_f).$$

□

LEMMA 4.3. *As $\eta\to 0$ and for $\theta\in[\nu_\eta+a,\lambda_e+a)$,*

$$\rho_{B,2}(\theta)\ge 2(1+r)\,\Gamma(\theta)\,q(\theta)-2r\log(K+1) \quad \textit{where,}$$
$$\Gamma(\theta)=\max_{i=0}^{K}\left(\mu_i\theta-\frac{1}{2}\mu_i^2-a\mu_i-\frac{1}{2}\lambda_f^2\right) \quad \textit{and}$$
$$q(\theta)=\Phi(\lambda_e-\theta)-\widetilde{\Phi}(a)-\widetilde{\Phi}(ar^{-1/2}).$$

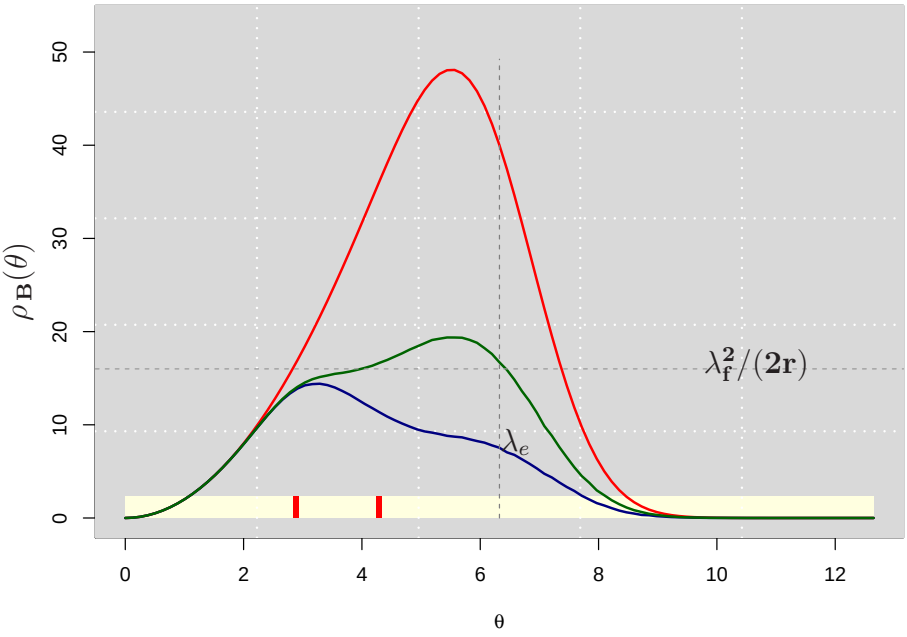

FIG 7. *Plot of the dominant portion of the predictive risk $\rho_D(\theta)$ as θ varies over the positive axis. In red, green and blue are respectively the risks of the optimal hard threshold plug-in estimator, unshared prediction scheme $\widehat{p}[r, T, \pi[\eta, r, 2], U]$ and the minimax optimal density estimate $\widehat{p}[r, T, CL^+, U]$. Here, $r = 0.25$, $\eta = e^{-20}$, $\lambda_f = 2.83$ and $\lambda_e = 6.32$. The red boxes at 2.83 and 4.24 in the yellow bar zone show the non-zero support point of the cluster prior $\pi[\eta, r, CL^+]$.*

In particular when $\theta \in (\lambda, \lambda + a)$, the bound shown in the Lemma can be negative which certainly proves that its crudeness as we already know that $\rho_{B,2}$ is always non-negative. However, the bound is intentionally kept crude as it helped to increase clarity in some of the later proofs.

PROOF. Using Inequality (22d) and the fact that $N_\theta(X, Y)$ is non-negative, we get the following lower bound:

$$\begin{aligned}
&\mathbb{E}_0\Big\{ N_\theta(X,Y)\, \mathbb{I}[X \le \lambda_e - \theta] \Big\} - \log(K+1) \\
&\ge v_w^{-1}\, \Gamma(\theta) \;\times\; P_0\big(X \le \lambda_e - \theta \text{ and } v_w^{-1}\, \mu_i(W+a) \ge 0 \text{ for } i = 0, 1, \ldots, K\big) \\
&= v_w^{-1}\, \Gamma(\theta) \;\times\; P_0\big(X \le \lambda_e - \theta \text{ and } W \ge -a\big)
\end{aligned}$$

as each μ_i is positive and we the bound the probability by

$$\begin{aligned}
&P_0\big(X \le \lambda_e - \theta,\ W \ge -a\big)\\
&\ge P_0\big(-a \le X \le \lambda_e - \theta \text{ and } X + Yr^{-1} \ge -a(1+r^{-1})\big)\\
&\ge P_0\big(-a \le X \le \lambda_e - \theta \text{ and } Y \ge -a\big)\\
&= P_0\big(-a \le X \le \lambda_e - \theta\big)\, P_0\big(Y \ge -a\big)\\
&= \big(\Phi(\lambda_e - \theta) - \widetilde{\Phi}(a)\big) \cdot \big(1 - \widetilde{\Phi}(ar^{-1/2})\big)\\
&\ge \Phi(\lambda_e - \theta) - \widetilde{\Phi}(a) - \widetilde{\Phi}(ar^{-1/2}).
\end{aligned}$$

Now, noting that $\rho_{B,2}(\theta) = 2r\mathbb{E}_0\Big\{N_\theta(X,Y)\,\mathbb{I}[X \le \lambda_e - \theta]\Big\}$ the result follows. □

LEMMA 4.4. *For* $\theta \in [\lambda_f + a, \lambda_e + a)$, $\quad \rho_{B,1}(\theta) - \rho_{B,2}(\theta) \le \lambda_f^2\,(1 + o\,(1))$.

PROOF. From Lemma 4.3, $\rho_{B,1}(\theta) - \rho_{B,2}(\theta)$ equals

$$\theta^2\left\{\widetilde{\Phi}(a) + \widetilde{\Phi}(ar^{-1/2})\right\} + \left\{\theta^2 - 2(1+r)\,\Gamma(\theta)\right\} q(\theta) + 2r\log(K+1).$$

In the similar way as in Lemma 4.2, we can show that as $\eta \to 0$, for all $\theta < \lambda_e + a$, $\theta^2\widetilde{\Phi}(a) = O(\lambda_f)$ and $\theta^2\widetilde{\Phi}(ar^{-1/2}) = o(\lambda_f^2)$.
And we show that the second sum involving $\theta^2 - 2(1+r)\,\Gamma(\theta)$ is bounded by $\lambda_f^2(1 + o(1))$ when $\theta \in [\lambda_f + a, \lambda_e + a)$. For this purpose, note that

$$\begin{aligned}
\theta^2 - 2(1+r)\Gamma(\theta) - \lambda_f^2 &= \min_{i=0}^{K}\left\{f_i(\theta) + 2(1+r)a\mu_i\right\}\\
&\le 2(1+r)\,a\,(\lambda_e + a) + \min_{i=0}^{K} f_i(\theta)\\
&\quad \text{where } f_i(\theta) = \theta^2 - 2(1+r)\mu_i\theta + (1+r)\mu_i^2 + r\lambda_f^2.
\end{aligned}$$

By construction of the cluster prior $\pi[\eta, K, r]$ the points μ_i were geometrically starting from $\mu_0 = \nu_\eta$ and with $\mu_{i+1} = (1+2r)\mu_i$ for all $i \in \{0, \ldots, K-1\}$. Also, the points end before $\lambda_e + a$. We have not used the properties of aligning rule anywhere before in our proof. A discrete, equiprobable prior distribute was all we utilized in the proofs before this stage.

Now, we will use the properties of μ_i. Note that for each $i \in \{0, \ldots, K\}$:

- f_i is convex in θ.
- $f_i(\mu_i) = \mu_i^2 - 2(1+r)\mu_i^2 + (1+r)\mu_i^2 + r\lambda_f^2 = -r(\mu_i^2 - \lambda_f^2) \le -r(\nu_\eta^2 - \lambda_f^2)$ as μ_i are increasing.

- Define $\mu_{K+1} = (1+2r)\mu_K$. Then by choice of $K, \mu_{K+1} > \lambda_e + a$. Also,

$$\begin{aligned} f_i(\mu_{i+1}) &= \mu_{i+1}^2 - 2(1+r)\mu_i\mu_{i+1} + (1+r)\mu_i^2 + r\lambda_f^2 \\ &= (\mu_{i+1} - (1+r)\mu_i)^2 - (1+r)r\mu_i^2 + r\lambda_f^2 \end{aligned}$$

and using the common ration of the geometric progression, we have

$$f_i(\mu_{i+1}) = r^2\mu_i^2 - (1+r)r\mu_i^2 + r\lambda_f^2 = -r(\mu_i^2 - \lambda_f^2) \leq -r(\nu_\eta^2 - \lambda_f^2).$$

Convexity of f_i implies that if $\mu_i \leq \theta \leq \mu_{i+1}$ for some $i \in \{1, \ldots, K\}$, then $f_i(\theta) \leq O(\lambda_f^2 - \nu_\eta^2) = o(\lambda_f^2)$ as by Equation [16], $\nu_\eta^2 - \lambda_f^2 \leq 2v_w^{1/2} a\lambda_f$. Hence, for all $\theta \in [\lambda_f + a, \lambda_e + a]$ we have $\min_{j=0}^K f_j(\theta) = o(\lambda_f^2)$. This complete the proof. □

5. Multivariate predictive risk.

In this section, we would need to construct sequence of priors as (n, s) varies. For notational convenience we assume s as a function of n here. It can easily be generalized. We consider a tractable convex collection of probability measures in the n–dimensional space

$$\mathcal{M}(n, s_n) = \left\{ \pi(\boldsymbol{\theta}) : \sum_{i=1}^n P_\pi(\theta_i \neq 0) \leq s_n \right\}.$$

However, $\mathcal{M}(n_n, s)$ contains prior whose support is not confined to $\Theta(n, s_n)$. We consider the sub-class $\mathcal{M}_p(n, s_n)$ of all product priors in $\mathcal{M}(n, s_n)$. The least favorable prior in $\mathcal{M}_p(n, s_n)$ concentrates on $\Theta(n, s_n)$ when $s_n \to \infty$ and $s_n/n \to 0$ as $n \to \infty$.

LEMMA 5.1. *For any n, s and r we have $R(n, s, r) \leq n\beta(s/n, r)$.*

PROOF. The set $\mathcal{M}(n, s)$ contains all Dirac priors $\delta_\theta \, \forall \, \theta \in \Theta(n, s)$ and is convex and weakly compact . So we can apply the Minimax Theorem 2.2 to have ,

$$R(n, s, r) \leq \sup\{B(\pi) : \pi \in \mathcal{M}(n, s)\} := B(\mathcal{M}(n, s), r)$$

and the result follows by Lemma A.3. □

Based on the univariate least favorable 3–point prior, for each $\epsilon < 1$, we construct a sequence (in n) of prior $\pi[n, s_n, \epsilon, r]$ in $\mathcal{M}^p(\epsilon s_n, r)$ as

$$\pi[n, s_n, \epsilon, r](\boldsymbol{\theta}) = \prod_{i=1}^n \pi[\epsilon s/n, r, 3](\theta_i).$$

The extension into the multivariate Bayes-Minimax set up can be conducted through the following lemma which is proved in the Appendix.

LEMMA 5.2. *As $n \to \infty$, $s \to \infty$ then if for each $\epsilon < 1$ there exists an exchangeable product prior $\pi_{n,\epsilon}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi_1[s/n, r](\theta_i)$ in $\mathcal{M}(n, s_n)$ satisfying the following conditions:*

a. $B(\pi_{n,\epsilon}) \geq \epsilon B(r, \mathcal{M}(n, \epsilon\, s_n))$

b. $\pi_{n,\epsilon}(\Theta(n, s)) \to 1$

c. $\int_{\Theta^c(n,s)} \pi_{n,\epsilon}(\boldsymbol{\theta})\, \boldsymbol{\rho}(\boldsymbol{\theta}, \widehat{p}_{t_{n,\epsilon}})\, d\boldsymbol{\theta} = o\left(B(r, \mathcal{M}(n, s_n))\right)$ *where* $t_{n,\epsilon} = \pi_{n,\epsilon}(\cdot|\Theta(n, s))$.

Then,

$$R(n, s, r) \sim B(r, \mathcal{M}(n, s_n)).$$

PROOF OF THEOREM 1.2. We check the conditions of the lemma for our least favorable prior. Consider the random variable N_n which is the number of non-zero coordinates in a random sample from $\pi[n, s_n, \epsilon, r]$. So, $N_n \sim \text{Binomial}(n, s_n/n)$.

As $s/n \to 0$ the 3-point prior is least favorable in $\mathfrak{m}(\epsilon s/n)$ and hence we have property (a) and Lemma A.3 implies $B(\pi[n, s_n, \epsilon, r]) \geq \epsilon\beta(\epsilon s_n, r)$. Property (b) also holds as

$$\pi[n, s_n, \epsilon, r]\big(\Theta^c(n, \epsilon s_n)\big) = P\left(N_n \geq \epsilon s_n\right) \leq \frac{\mathbb{V}ar(N_n)}{(1-\epsilon)^2 \mathbb{E}^2(N_n)}$$

which by Chebyshev's inequality goes to 0 as $s_n \to \infty$.

Now, note that the support of $\pi[n, s_n, \epsilon, r]$ is given by

$$S_{n,\epsilon} = \left\{\zeta : \zeta_i = 0 \text{ or } \pm\nu_\eta \text{ and } N_n(\zeta) \leq s_n\right\}$$

where ν_η is given by Equation [16] with $\eta = n^{-1}\epsilon s_n$. And, the univariate plug-in risk $\rho(\theta, \widehat{p}_E[0]) = \theta^2/(2r)$ and $\rho(\theta, \widehat{p}_E[\pm\nu_\eta]) = (\theta \pm \nu_\eta)^2/(2r)$.

So, by convexity of the relative entropy loss function we have,

$$\begin{aligned}
\boldsymbol{\rho}\left(\boldsymbol{\theta},\, \widehat{p}\big[\pi[n, s_n, \epsilon, r]\,\big]\right) &\leq \sup_{\zeta \in S_{n,\epsilon}} \boldsymbol{\rho}\left(\boldsymbol{\theta},\, \widehat{p}_E[\zeta]\right) \\
&\leq \frac{1}{2r}\left[\sum_{i:\zeta_i=0} \theta_i^2 + \sum_{i:\zeta_i \neq 0} (\theta_i \pm \nu_\eta)^2\right] \\
&\leq r^{-1}\left\{||\boldsymbol{\theta}||_2^2 + N_n\, \nu_\eta^2\right\} = 2r^{-1} N_n \nu_\eta^2.
\end{aligned}$$

Now integrating over the prior π, we have

$$\int_{\boldsymbol{\theta}\in\Theta^c(n,s_n)} \pi[n,s_n,\epsilon,r](\boldsymbol{\theta})\ \boldsymbol{\rho}\left(\boldsymbol{\theta},\widehat{p}\big[\pi[n,s_n,\epsilon,r]\,\big]\right) d\boldsymbol{\theta} = 2r^{-1}\nu_\eta^2 E\{N_n;\,\Theta_n^c\}.$$

Extending the univariate minimax problem it follows that $B(r,\mathcal{M}(n,s_n)) \sim (2r)^{-1}s_n\nu_\eta^2$ and so the ratio

$$\big\{B(r,\mathcal{M}(n,s_n))\big\}^{-1}\left\{\int_{\boldsymbol{\theta}\in\Theta^c(n,s_n)} \pi[n,s_n,\epsilon,r](\boldsymbol{\theta})\,\boldsymbol{\rho}\big(\boldsymbol{\theta},\widehat{p}[\,\pi[n,s_n,\epsilon,r]\,]\big)\,d\boldsymbol{\theta}\right\}$$

is asymptotically equal to $\mathbb{E}[N_n\mathbb{I}\{\Theta^c(n,\epsilon s_n)\}]/\mathbb{E}(N_n)$, which converges to 0 as $n\to\infty$. The convergence is a consequence of the concentration of N_n as $N_n(\theta) = \mathbb{E}N_n\ (1+o(1))$ which follows directly from Chebyshev's inequality.

Thus, property (c) of the Lemma is also satisfied and $\prod_{i=1}^n \pi[n,s_n/n,r,3](\theta_i)$ is the asymptotically least favorable prior and the theorem follows. □

6. Further Insights into Minimax strategies.

6.1. *The choice of Threshold.* For having an optimal threshold density estimate we need to have a minimum threshold of λ_e. The working principle of threshold rule is that after the threshold zone they use an estimator with bounded bias and as a side-effect it has considerable loss at the origin. So, it needs ideal calibration of the threshold as the higher thresholds decreases the risk contribution from above the threshold ρ_A at the origin.

Based on the calculations in Section 4 it follows that we need to restrict $\rho_A(0)$ below the minimax risk $\eta\lambda_f^2/2r$. If the threshold is $t = q\lambda_e$, where $0 < q < 1$, then from the calculations in Equation [23] we have

$$\rho_A(0) = O\big(t\,\phi(t)\big) = O\big(te^{-t^2/2}\big) = O\big(\eta^{q^2}\big) >> O(\eta\lambda_f^2).$$

So, λ_e is the minimal threshold that is to be used. This proves Lemma 1.3. Note that the fact that the threshold is controlled entirely by the degree of sparsity resonates with general philosophy of this sparse predictive regime where the order of the optimal risk depends on sparsity and is scaled by future uncertainty.

6.2. *Sub-optimality of $\mathcal{L},\mathcal{E}$ and $\mathcal{G}$.* Based on the minimax risk of sparse estimation of the normal mean and the calculations in the Appendix A.1.2, we see that Plug-in density estimates and in general the class of Gaussian predictive densities is minimax sub-optimal with the sub-optimality ratio being independent of η.

LEMMA 6.1. *For any fixed $r \in (0,\infty)$, under the condition of Theorem 1.2, we have*

$$R(n,s,r,\mathcal{E}) \sim (1+r^{-1})\, R(n,s,r).$$

Also, minimax optimal density estimates lie outside $\mathcal{G}$. Lemma 1.2 follows by using Theorem 1.2 and the following lemma from Mukherjee and Johnstone (2012b).

LEMMA 6.2. *In model M.2 as $n \to \infty$ and $s/n \to 0$ we have*

$$\liminf_{n\to\infty} \min_{\widehat{p}\in\mathcal{G}_n} \max_{\boldsymbol{\theta}\in\Theta_n(s)} \rho\,(\boldsymbol{\theta},\widehat{p}) \;\geq\; (1+\boldsymbol{r})^{-1}\, s\, \log(n/s)\big(1+o\,(1)\big).$$

The sub-optimality of Linear density estimates as described in Lemma 1.1 follows from the risk calculations in Appendix A.1.3.

6.3. *Risk sharing schemes and efficient alignments of support points.* While constructing the minimax threshold estimator $\widehat{p}_T$ we have used the cluster prior $\pi[\eta,r,K]$ below the threshold. This choice is not unique and we can use other proper estimates in its place. By the proof in the Section 4, it follows that we can not use zero-threshold estimator as then ρ_B will only require $\rho_{B,1}$ and for optimality we also need the sharing effect from $\rho_{B,2}$. The proof structure outline before Lemma 4.4 goes through for any finite prior with $(1-\eta)$ probability at the origin and the remaining probability η being equally allocated across finite points between λ_f and λ_e.
So a prior with $(1-\eta)$ mass at the orgin and sharing the remaining mass η across the non-zero support points $\mu_0,\ldots,\mu_{K_1}$ will produce an optimal allocation if the alignment of these points (and the cardinality of the support) is such that Lemma 4.4 still holds.

For example, instead of the cluster prior $\pi[\eta,r,K]$ we choose a (K_1+2)-points prior whose non-zero support points are equispaced (unlike in geometric progression) and equiprobable. Let the spacing between the points be s. So, $\mu_0=\nu_\eta$ and $\mu_i=\mu_0+i\,s$ for $i\in\{0,\ldots,K_1\}$ where

$$K_1 = \max\{i : \mu_0 + i\,s \leq \lambda_e + a\} \text{ and so as } \eta\to 0,\; K_1 \sim \left\lfloor \frac{\lambda_e-\lambda_f}{s} \right\rfloor.$$

Again if we would like to equate $f_i(\mu_{i+1}) \leq -r(\mu_i^2-\lambda_f^2)$ as in Lemma 4.4 we have,

$$\begin{aligned} f_i(\mu_{i+1}) &= (\mu_{i+1}-(1+r)\mu_i)^2 - (1+r)r\mu_i^2 + r\lambda_f^2 \\ &= (s-r\mu_i)^2 - (1+r)r\mu_i^2 + r\lambda_f^2 \\ &= -r\mu_i^2 + r\lambda_f^2 + s^2 - 2sr\mu_i. \end{aligned}$$

Now, we solve for the condition $s\mu_0(s\mu_0 - 2r\mu_i) \leq 0$. A solution is given by $s = 2r\lambda_f$ which produces a choice of $K_1 = \lfloor (2r)^{-1}\{(1+r^{-1})^{1/2} - 1\} \rfloor$.

We construct the following univariate prior with the non-zero support , equiprobable and equidistant support points lying in between μ_0 and $\lambda + a$

$$\pi[\eta, r, E]\theta) = (1-\eta)\cdot\delta_0(\theta) + \frac{\eta}{K_1+1}\sum_{i=0}^{K_1}\delta_{\mu_i}(\theta)$$

$$\text{where } \mu_i = (1+2ri)\mu_0,\ i = 1,\ldots,K_1 \ \text{ and } \ K_1 = \left\lfloor \frac{(1+r^{-1})^{1/2}-1}{2r} \right\rfloor .$$

As $\eta \to 0$, it will produce a first order minimax optimal predictive density estimate for the univariate restricted Bayes-Minimax problem over the constrained prior space $\mathfrak{m}^+(\eta)$.

In Figure [8] we have the empirical evaluations of optimal predictive schemes in the asymptotic regime. Figure [9] contains the risk plots under moderate sparsity.

6.4. *Other Minimax Estimators.*

We consider the following non-negative analogue of $\pi[\eta, r, \text{INF}]$

$$\pi[\eta, r, \text{INF}^+](\theta) = (1-\eta)\cdot\delta_0(\theta) + (1-\eta)\sum_{j=0}^{\infty}\eta^{j+1}\sum_{i=0}^{K} s_i\,\delta_{\mu_{ij}}(\theta) \quad \text{where,}$$

$$\mu_{ij} = j\,\lambda_e + (1+2r)^i\,\nu_\eta\,;\ i = 0,\ldots,K \text{ and } j = 1,\ldots,\infty$$

$$s_i = (\log\eta^{-1})^{-i} \text{ for } i = 1,\ldots,K \text{ and}$$

$$s_0 = 1 - \frac{(1-(\log\eta^{-1})^{-K})}{(\log\eta^{-1}-1)} \sim 1-(\log\eta^{-1})^{-1} \text{ as } \eta \to 0.$$

The between cluster spacing and probability distribution on the clusters is motivated by the construction of second order minimax optimal point estimates of the normal mean in Johnstone (1994). The with-in cluster mass distribution is more interesting. First note that for the univariate predictive density estimation problem $\pi[\eta, r, 2]$ or its corresponding infinite support geometric version is not the Bayes optimal strategy for the statistician because it is a prediction problem and he has to share his future risk. Also, neither $\pi[\eta, r, \text{CL}^+]$ nor its corresponding infinite support geometric version is least favorable as after sharing the statistician incurs the maximum risk at λ_f (which is expected) and the risks at the other non-zero support points

is appreciably lower even in first order calculations. However, as $\eta \to 0$ a geometrically decreasing discrete probability sharing scheme with common ratio $\log \eta^{-1}$ solves this problem because $\log \eta^{-1} \to \infty$ when $\eta \to 0$ and hence dampens the first-order terms in the asymptotic limit.

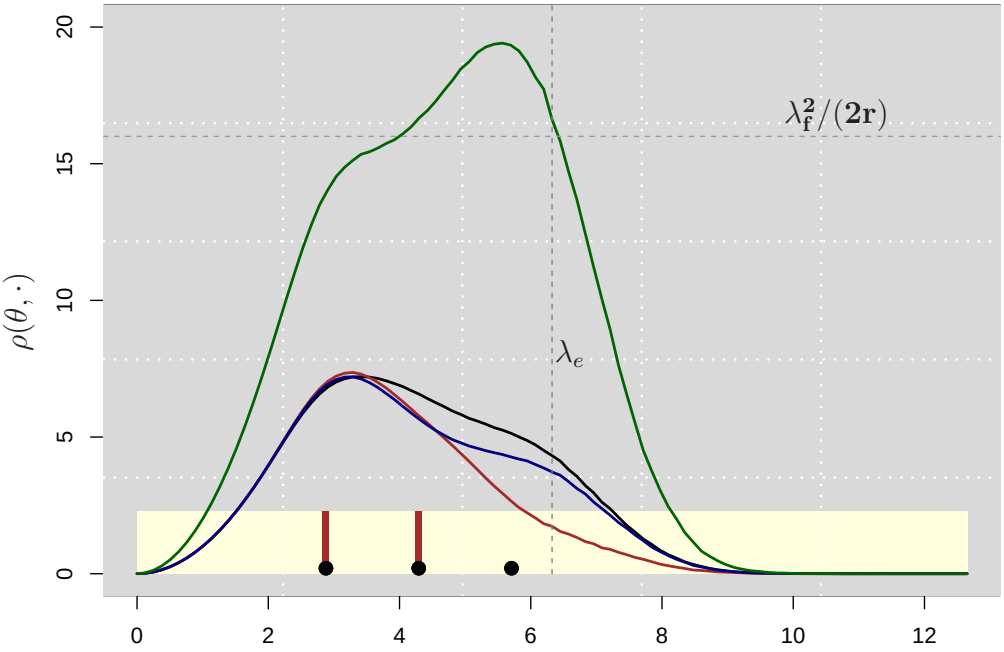

FIG 8. *Plot of the predictive entropy risk $\rho(\theta, \cdot)$ for the different univariate predictive schemes as the parameter θ varies over $\mathbb{R}^+$. In green, blue, brown and black are respectively the risks of $\widehat{p}[r, T, \pi[\eta, r, 2], U]$, $\widehat{p}[r, T, CL^+, U]$, $\widehat{p}[r, T, \pi[\eta, r, E], U]$ and that of the Bayes predictive density estimate based on the infinite support prior $\pi[\eta, r, INF]$. Here, $r = 0.25$, $\eta = e^{-20}$, $\lambda_f = 2.83$ and $\lambda_e = 6.32$. The brown boxes at 2.83 and 4.24 in the yellow zone show the non-zero support point of the cluster prior $\pi[\eta, r, CL^+]$ and the black circles denote the non-zero support points of $\pi[\eta, r, E]$.*

PROOF OF THEOREM 1.3. We prove the result in the corresponding univariate model $\mathbf{M.2}(1, \eta, r)$ with $\sigma_p = 1$ and the parameter space restricted to the non-negative axis. The risk of $\widehat{p}(y|x, \pi[\eta, r, \text{INF}^+])$ at the point θ is given by:

$$\rho(\theta) = \frac{1}{2r}\left(\theta^2 - 2rE_0(N'_\theta(X, Y)) + 2rE_0(D'_\theta(X))\right) \quad \text{where,}$$

$$N'_\theta = \log\left[1+\sum_{j=0}^{\infty}\frac{\eta^{j+1}}{K+1}\sum_{i=0}^{K}\exp\left\{\mu_{ij}\left(x+\frac{y}{r}\right)-\frac{1+r}{2r}\mu_{ij}^2+\frac{1+r}{r}\mu_{ij}\theta\right\}\right]$$

$$D'_\theta = \log\left[1+\sum_{j=0}^{\infty}\frac{\eta^{j+1}}{K+1}\sum_{i=0}^{K}\exp\{\mu_{ij}x-\frac{1}{2}\mu_{ij}^2+\mu_{ij}\theta\}\right].$$

Now, $\rho(0)=\eta+o(\eta)$ as $N'_\theta(X,Y)\geq 0$ and

$$E_0(D'_0(X)) \leq \log\left[1+\sum_{j=0}^{\infty}\sum_{i=1}^{K}\frac{\eta^{j+1}}{K+1}\right] = -\log(1-\eta)=\eta+O(\eta^2).$$

Next we note that, with probability 1 we will have,

$$E_0(N'_\theta(X,Y)) \geq \max_{\substack{i=1,\ldots,K\\ j=0,\ldots,\infty}}\left(\mu_{ij}\theta-\frac{1}{2}\mu_{ij}^2-a\mu_{ij}-\frac{1}{2}(j+1)\mu_0^2\right)\left(1+\frac{1}{r}\right)$$
$$-\log(K+1)+\mathrm{O}(1) \quad \text{and}$$
$$E_0(D'_\theta(X)) \leq \max_{\substack{i=1,\ldots,K\\ j=0,\ldots,\infty}}\left(\mu_{ij}\theta-\frac{1}{2}\mu_{ij}^2+a\mu_{ij}-\frac{1}{2}(j+1)\lambda^2\right)_+ +\mathrm{O}(1).$$

Also, the optimum value of j and i for both the numerator and denominator is same. And so it follows

$$\rho(\theta) \leq \frac{1}{2r}\max_{\substack{i=1,\ldots,K\\ j=0,\ldots,\infty}}(\theta^2-2\mu_{ij}\theta+\mu_{ij}^2+4a\mu_{ij})+\mathrm{O}(1)$$
$$\leq \frac{1}{2r}\max_{\substack{i=1,\ldots,K\\ j=0,\ldots,\infty}}(\theta^2-2\mu_{ij})^2+o(\mu_0^2) \leq \mu_0/2r(1+\mathrm{o}(1)).$$

Risk calculations similar to Section 3 will show that the $B(\pi[\eta,r,\mathrm{INF}^+])$ attains the above lower bound. This will complete the proof. □

7. Discussion. The decision theoretic techniques used here can describe the operating characteristics of a large class of predictive strategies. The predictive minimax theory developed here can be extended to weak ℓ_p balls of varying shape and sizes (Mukherjee and Johnstone, 2012a). The calculations here are first order. Second order calculations would reflect further relations between estimation and prediction theory. In the non-orthogonal model instead of r, the minimax risk would depend on the eigen structure of $B'B(A'A)^{-1}$.

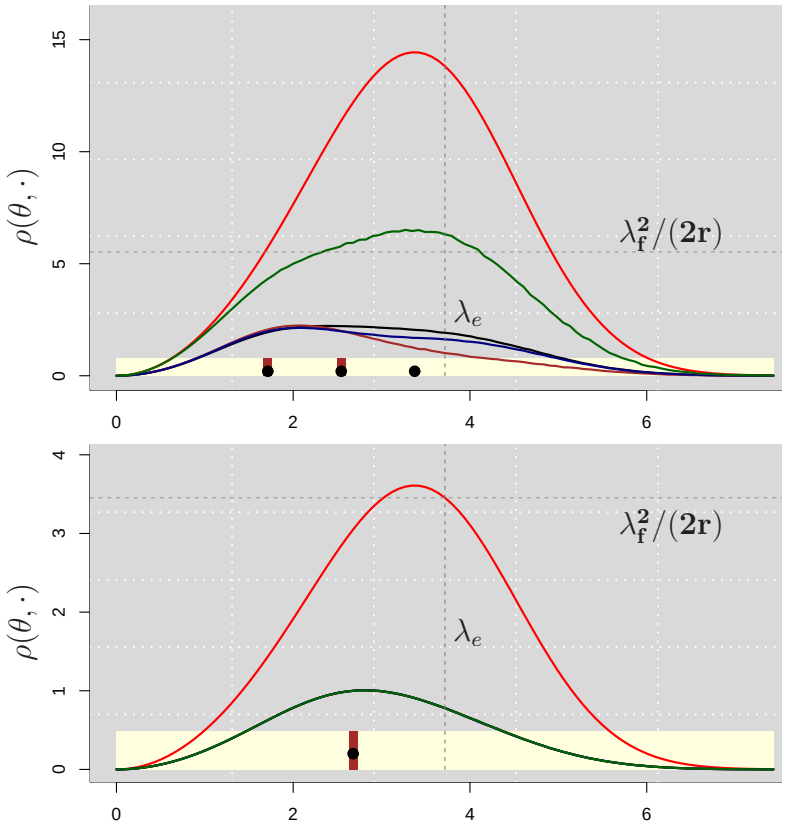

FIG 9. *The figure shows the risk plots for the different univariate predictive schemes under moderate degree of sparsity* ($\eta = 0.001$) *for the two different values of the future to past variances:* $r = 0.25$ *(top) and* $r = 1$. *In red, green, blue, brown and black are respectively the risks of the optimal hard-threshold plug-in scheme,* $\widehat{p}[\,r,\, T, \pi[\eta, r, 2],\, U]$, $\widehat{p}[\,r,\, T,\, CL^{+},\, U]$, $\widehat{p}[\,r,\, T, \pi[\eta, r, E],\, U]$ *and that of the Bayes predictive density estimate based on the infinite support prior* $\pi[\eta, r, INF]$.

APPENDIX A

A.1. Basic Decision Theory Results in M.2.

As the true density $\phi(\cdot|\boldsymbol{\theta}, \sigma_f^2)$ in $\mathbf{M.2}(n, s, \sigma_e, \sigma_f)$ is bounded above by $c_n = (2\pi\sigma_f^2)^{-n/2}$, we can restrict the action set $\mathcal{A}_n$ comprising of all densities in $\mathbb{R}^n$ to the set of all c_n bounded densities

$$\mathcal{A}(n, c_n) = \left\{ p : \mathbb{R}^n \to \mathbb{R} \text{ such that } \int_{\mathbb{R}^n} p\,(\mathbf{y})\, d\mathbf{y} = 1 \text{ and } p \in [0, c_n] \right\}.$$

LEMMA A.1. *For any* $p \in \mathcal{A}_n$ *but not in* $\mathcal{A}(n, c_n)$ *there exists* $p_b \in \mathcal{A}(n, c_n)$ *that dominates* p *in the sense* $\boldsymbol{L}(\boldsymbol{\theta}, p_b) \leq \boldsymbol{L}(\boldsymbol{\theta}, p)$ *for all* $\boldsymbol{\theta} \in \mathbb{R}^n$.

Details of the proof can be found in Brown, George and Xu (2008, Lemma2). For any positive density $p \in \mathcal{A}_n$ but not in $\mathcal{A}(n, c_n)$ the general idea for constructing a better estimate $p_b \in \mathcal{A}(n, c_n)$ is to truncate p on the set $S_p = \{\mathbf{y} \in \mathbb{R}^n : p(\mathbf{y}) > c_n\}$ and to lift it on S_p^c, i.e.

$$p_b(\mathbf{y}) = \begin{cases} \{\int_{S_p^c} p(\mathbf{y})\, d\mathbf{y}\}^{-1} \cdot \{1 - c_n \mathrm{Vol}(S_p)\} \cdot p(\mathbf{y}) & \text{if } \mathbf{y} \in S_p^c \\ c_n & \text{if } \mathbf{y} \in S_p \end{cases}.$$

As the KL loss for estimators in $\mathcal{A}(n, c_n)$ is always defined (can be infinite though), they are mathematically more convenient for risk calculations than estimator in $\mathcal{A}_n \setminus \mathcal{A}(n, c_n)$. In the light of Lemma A.1, with out any loss of generality we restrict ourselves to estimators in $\mathcal{A}(n, c_n)$ only. Next we show that the Bayes predictive density defined in equation (4) actually minimizes the integrated Bayes risk for any prior π in collection $\mathcal{P}(\mathbb{R}^n)$ of all probability measures on $\mathbb{R}^n$.

LEMMA A.2. *For any prior $\pi \in \mathcal{P}(\mathbb{R}^n)$ if $B(\pi, \widehat{p}_\pi) < \infty$ then we have*

$$B(\pi, \widehat{p}_\pi) \leq B(\pi, \widehat{p}) \text{ for any } \widehat{p} \in \mathcal{A}_n.$$

PROOF. As the true density $\phi(\cdot|\boldsymbol{\theta}, \sigma_f^2)$ is bounded above, the marginal density $m_\pi(\mathbf{x}) < \infty$ almost surely for all $\mathbf{x} \in \mathbb{R}^n$. Thus, $\widehat{p}_\pi$ is defined almost everywhere and by Lemma A.1, with out loss of generality we can assume that $\widehat{p} \in \mathcal{A}(n, c_n)$. The difference in the integrated Bayes risk between any estimator $\widehat{p}$ and the Bayes estimator is,

$$B(\pi, \widehat{p}) - B(\pi, \widehat{p}_\pi) = \iiint \phi(\mathbf{x}|\boldsymbol{\theta}, \sigma_p^2)\, \phi(\mathbf{y}|\boldsymbol{\theta}, \sigma_f^2)\, \pi(\boldsymbol{\theta}) \log \frac{\widehat{p}_\pi(\mathbf{y}|\mathbf{x})}{\widehat{p}(\mathbf{y}|\mathbf{x})}\, d\boldsymbol{\theta}\, d\mathbf{y}\, d\mathbf{x}$$

as we can interchange the order of integrals by Fubini's theorem. Also, $m_\pi(\mathbf{x})\, \widehat{p}_\pi(\mathbf{y}|\mathbf{x}) = \int \phi(\mathbf{x}|\boldsymbol{\theta}, \sigma_p^2)\, \phi(y|\boldsymbol{\theta}, \sigma_f^2)\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$, so we have,

$$B(\pi, \widehat{p}) - B(\pi, \widehat{p}_\pi) = \iint m_\pi(\mathbf{x})\, \widehat{p}_\pi(\mathbf{y}|\mathbf{x}) \log \frac{\widehat{p}_\pi(\mathbf{y}|\mathbf{x})}{\widehat{p}(\mathbf{y}|\mathbf{x})}\, d\mathbf{y}\, d\mathbf{x}$$

which is the KL divergence between the densities $m_\pi(\mathbf{x}) \times \widehat{p}_\pi(\mathbf{y}|\mathbf{x})$ and $m_\pi(\mathbf{x}) \times \widehat{p}(\mathbf{y}|\mathbf{x})$ and so it always non-negative. This completes the proof. □

A.1.1. *Minimax Theorem.*

We consider the Gaussian predictive sequence model

$$x_i = \theta_i + \sigma_p\, \epsilon_{1,i} \text{ and } y_i = \theta_i + \sigma_p\, \epsilon_{2,i} \tag{26}$$

for $i \in I \subset \mathbb{N}$, with $\epsilon_{1,i}$ and $\epsilon_{2,i}$ are i.i.d. $N(0,1)$ random variables. The parameter space is a collection of $\boldsymbol{\theta}$ for which $\sum_i \theta_i^2 < \infty$ and is denoted by $\ell_2(\mathbb{N})$. The action set is given by

$$\mathcal{A}_\infty = \left\{ p : \mathbb{R}^\infty \to \mathbb{R} \text{ such that } \int_{\mathbb{R}^\infty} p(\mathbf{y})\, d\mathbf{y} = 1 \text{ and } p(\mathbf{y}) \geq 0 \text{ for all } \mathbf{y} \right\}.$$

For each $n \in \mathbb{N}$ consider all sub-probabilities in $\mathbb{R}^n$ bounded by c_n by extending $\mathcal{A}(n, c_n)$ to its closure

$$\bar{\mathcal{A}}(n, c_n) = \left\{ p : \mathbb{R}^n \to \mathbb{R} \text{ such that } \int_{\mathbb{R}^n} p\,(\mathbf{y})\, d\mathbf{y} \leq 1 \text{ and } p \in [0, c_n] \right\}.$$

$\bar{\mathcal{A}}(n, c_n)$ is a sub-set of the Banach space $\mathcal{L}_\infty(\mathbb{R}^n, \mathbb{R})$ – all bounded functionals in $\mathbb{R}^n$. We consider the topology on $\bar{\mathcal{A}}(n, c_n)$ induced by the weak* topology on $\mathcal{L}_\infty(\mathbb{R}^n, \mathbb{R})$.

In the predictive sequence model consider the experiment $(\Omega, \mathcal{B}, \{\mathcal{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \ell_2(\mathbb{N})\})$ where the sample space $\Omega = \{\otimes_{i \in \mathbb{N}}(x_i, y_i) : x_i, y_i \in \mathbb{R}\}$ and $\mathcal{B}$ is the associated Borel sigma field. As the parameter space is ℓ_2, we have a dominated experiment here with

$$\frac{d\mathcal{P}_{\boldsymbol{\theta}}}{d\mathcal{P}_0} = \exp\left[\sigma_p^{-2} \left\{ \left\langle \boldsymbol{\theta}, \mathbf{x} + r^{-1}\mathbf{y} \right\rangle - \frac{1}{2}\left(1 + r^{-1}\right) ||\boldsymbol{\theta}||^2 \right) \right\} \right].$$

Also, for each $n \in \mathbb{N}$, the restricted, closed action set $\bar{\mathcal{A}}(n, c_n)$ is weak* compact and the loss function $\mathbf{L}(\boldsymbol{\theta}, p)$ is

- strictly convex in $p \in \bar{\mathcal{A}}(n, c_n, +)$ for any $\boldsymbol{\theta} \in \mathbb{R}^n$ where $\bar{\mathcal{A}}(n, c_n, +) = \{p \in \bar{\mathcal{A}}(n, c_n) \text{ such that } \mathbf{L}(\boldsymbol{\theta}, p) < \infty \text{ for all } \boldsymbol{\theta} \in \mathbb{R}^n\}$ and
- lower semi-continuous in p on $\bar{\mathcal{A}}(n, c_n)$ for any fixed $\boldsymbol{\theta} \in \mathbb{R}^n$.

So, following the lines of Brown (1974) and Johnstone (2012, Appendix 1) the following minimax theorem can be attained in the predictive setting.

THEOREM A.1. *Consider the predictive density estimation problem in the Gaussian predictive sequence model (26) with the parameter space $\ell_2(\mathbb{N})$. For any convex set $\mathcal{M}$ of probability measures on $\ell_2(\mathbb{N})$ we have*

$$\inf_{\widehat{p} \in \mathcal{A}_\infty} \sup_{\pi \in \mathcal{M}} B\left(\pi, \widehat{p}\right) = \sup_{\pi \in \mathcal{M}} \inf_{\widehat{p} \in \mathcal{A}_\infty} B\left(\pi, \widehat{p}\right)$$

Next, we describe the predictive risk of the different sub-class of estimators in model **M.2** with $\sigma_p = 1$ and $\sigma_f = r$.

A.1.2. *Plug-in Risk.*

The risk of the plug-in predictive density $\widehat{p}_E[\,\widehat{\boldsymbol{\theta}}\,] = \phi(\mathbf{y}|\,\widehat{\boldsymbol{\theta}}\,(\mathbf{X})\,, r)$ is given by

$$\begin{aligned}\boldsymbol{\rho}\Big(\boldsymbol{\theta},\, \phi(\,\cdot\,|\,\widehat{\boldsymbol{\theta}}, r)\,\Big) &= \mathbb{E}_{\boldsymbol{\theta}}\left[\log\left(\frac{\phi(\mathbf{Y}|\,\boldsymbol{\theta}, r)}{\phi(\mathbf{Y}|\,\widehat{\boldsymbol{\theta}}, r)}\right)\right] \\ &= (2r)^{-1}\,\mathbb{E}_{\boldsymbol{\theta}}\left(-\|\mathbf{Y}-\boldsymbol{\theta}\|^2 + \|\mathbf{Y}-\widehat{\boldsymbol{\theta}}(\mathbf{X})\|^2\right)\end{aligned}$$

where the expectation is over $\phi\big(\mathbf{x}|\,\widehat{\boldsymbol{\theta}}, r\big)\,\times\,\phi\big(\mathbf{y}|\,\widehat{\boldsymbol{\theta}}, r\big)$ – the joint density of $(\mathbf{X}, \mathbf{Y})$ under the true location $\boldsymbol{\theta}$. On expanding the second term in the above sum, the crossproduct term will vanish due to time-invariance and so we have

$$\boldsymbol{\rho}\Big(\boldsymbol{\theta},\, \phi(\,\cdot\,|\,\widehat{\boldsymbol{\theta}}, r)\,\Big) = (2r)^{-1}\mathbb{E}_{\boldsymbol{\theta}}\|\widehat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\theta}\|^2 = (2r)^{-1}q(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}, 1) \tag{27}$$

Thus, the plug-in KL predictive risk is the quadratic location risk discounted by the future variability, and so the risk properties of plug-in density estimates follow directly from point estimation theory. Plug-in densities are also called estimative densities and the plug-in predictive risk will sometimes be also denoted by $\boldsymbol{\rho}_{\boldsymbol{E}}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}})$.

A.1.3. *Risk of Linear predictive density estimates.*

This class consists of Bayes density estimates based on conjugate product normal priors $\prod_{i=1}^{n}\phi(\,.\,|0, l_i)$ where $l_i, i = 1, 2, \ldots, n$ are the non-negative prior variances. They are referred to as "Linear" predictive densities because by Diaconis and Ylvisaker (1979) they are analogous to linear diagonal point estimates in the quadratic error setting.

We exhibit the calculation for the univariate case. As the normal prior $N(0, l)$ is conjugate, the posterior is also normal. Setting $\alpha = (1 + l^{-1})^{-1}$, we have

$$\pi(\theta\,|\,x) \propto \phi_1(x - \theta\,|\,0, 1)\;\times\;\phi_1(\theta\,|\,0, l) \propto \phi_1(\theta\,|\,\alpha x, \alpha) \sim N(\alpha x,\, \alpha)$$

The corresponding predictive density is a convolution of Gaussians

$$\widehat{p}_l(y|x) = \int \phi_1(y - \theta\,|\,0, r)\,\phi_1(\theta - \alpha x\,|\,0, \alpha)\,d\theta$$

and so is also Gaussian: $\widehat{p}_l(y|x) \sim N(\alpha\, x,\, r + \alpha)$.

And given the past x, its loss is given by:

$$\begin{aligned}
&\mathbf{L}\big(\theta, \widehat{p}_l\,(\,.\,|\,X=x)\big) = \mathbb{E}_\theta \log \phi(Y-\theta\,|\,0,\,r) - \mathbb{E}_\theta \log \phi(Y-\alpha x\,|\,0,\,r+\alpha) \\
&= -\frac{1}{2}\bigg(\log(2\pi r) + \frac{\mathbb{E}_\theta (Y-\theta)^2}{r}\bigg) + \frac{1}{2}\bigg(\log\big(2\pi(r+\alpha)\big) + \frac{\mathbb{E}_\theta\,(Y-\alpha x)^2}{2\,(r+\alpha)}\bigg) \\
&= \frac{1}{2}\log\bigg(1+\frac{\alpha}{r}\bigg) + \frac{\mathbb{E}_\theta\,(Y-\alpha x)^2}{2\,(r+\alpha)} - \frac{1}{2}
\end{aligned}$$

and by Bias-Variance decomposition, we have, $\mathbb{E}_\theta\,(Y-\alpha x)^2 = (\theta-\alpha x)^2 + r$. To evaluate the risk we take expectation over the past X and again by Bias-Variance decomposition $\mathbb{E}_\theta\,(Y-\alpha X)^2 = (1-\alpha)^2\,\theta^2 + \alpha^2 + r$. So, the risk is

$$\begin{aligned}
\rho\big(\,\theta, \widehat{p}_l\,\big) &= \frac{1}{2}\log\bigg(1+\frac{\alpha}{r}\bigg) + \frac{(1-\alpha)^2\theta^2 - \alpha(1-\alpha)}{2\,(r+\alpha)} \\
&= \frac{1}{2}\log\bigg(1+\frac{\alpha}{r}\bigg) + \frac{(1-\alpha)^2}{2\,(r+\alpha)}\,(\theta^2 - l).
\end{aligned}$$

As $l \to \infty, \quad \alpha \to 1$, we get the uniform prior Bayes Predictive Density $(\widehat{p}_U)$. It is the best invariant density as well as minimax in the unrestricted parametric space (will be referred as 'canonical minimax estimator'). Thus:

$$\widehat{p}_U\big(y\,|\,X=x\big) = N(x, 1+r) \quad \text{and} \quad r(\theta, \widehat{p}_U) = \frac{1}{2}\log\big(1+r^{-1}\big)\,.$$

Again, as $l \to 0, \quad \alpha \to 0$ and we get the zero density $\phi(\,\cdot\,|\,0, r)$ with $\theta^2/2r$ as its risk.

Returning back to the multivariate case, the KL risk of the linear predictive density estimate

$$\widehat{p}_L[\boldsymbol{\alpha}] = \prod_{i=1}^{n} \phi\big(\,.\,|\,\alpha_i\,X[i],\,(\alpha_i + r)\big), \alpha_i = (1+l_i^{-1})^{-1},\, l_i \geq 0,\; i = 1, \ldots, n$$

is quadratic in the parameter $\boldsymbol{\theta}$ and is given by

$$\boldsymbol{\rho}\big(\,\boldsymbol{\theta}, \widehat{p}_L[\boldsymbol{\alpha}]\,\big) = \frac{1}{2}\sum_{i=1}^{n}\log\bigg(1+\frac{\alpha_i}{r}\bigg) + \sum_{i=1}^{n}\frac{(1-\alpha_i)^2}{2\,(r+\alpha_i)}\,(\theta_i^2 - l_i).$$

A.1.4. *Gaussian Predictive Risk.*

The loss of the density estimate $\widehat{p}_G[\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{d}}](\mathbf{y}|\mathbf{X}) = \prod_{i=1}^n \phi(y_i|\widehat{\theta}_i(\mathbf{X}), \widehat{d}_i(\mathbf{X}))$ is given by,

$$\mathbf{L}\Big(\boldsymbol{\theta}, \widehat{p}_G\big[\,\widehat{\boldsymbol{\theta}},\, \widehat{\mathbf{d}}\,\big]\big(\,\cdot\,\big|\mathbf{X}=\mathbf{x}\big)\Big) = \sum_{i=1}^n \mathbb{E}_{\boldsymbol{\theta}} \log\left(\frac{\phi\big(Y_i\big|\,\theta_i, r\big)}{\phi\big(Y_i\big|\,\widehat{\theta}_i, \widehat{d}_i\,\big)}\right)$$
$$= \frac{1}{2}\sum_{i=1}^n \left[\log\left(r^{-1}\widehat{d}_i\right) + \frac{r + \|\widehat{\theta}_i - \theta_i\|^2}{\widehat{d}_i} - 1\right].$$

Integrating the loss over the past $\mathbf{X}$, we find the risk

$$\boldsymbol{\rho}\Big(\boldsymbol{\theta}, \widehat{p}_G\big[\,\widehat{\boldsymbol{\theta}},\, \widehat{\mathbf{d}}\,\big]\Big) = \frac{1}{2}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{\theta}}\left[\log\left(r^{-1}\widehat{d}_i(\mathbf{X})\right) + \frac{r + \|\widehat{\theta}_i(\mathbf{X}) - \theta_i\|^2}{\widehat{d}_i(\mathbf{X})} - 1\right].$$

A.1.5. *Mills ratio and Gaussian tails.*

The function $M(u) = \widetilde{\Phi}(u)/\phi(u)$ is called Mills Ratio. The following inequalities will provide an approximation to the Mills ratio which will be very helpful for our calculations with truncated Gaussian random variable. By Johnstone (2012, Exercise 8.1):

$$\text{for any } u \geq 0 \text{ we have } \frac{\phi(u)}{u}\left(1 - \frac{1}{u^2}\right) \leq \widetilde{\Phi}(u) \leq \frac{\phi(u)}{u}. \tag{28}$$

And so for large u, which will typically be the case, the approximation $\widetilde{\Phi}(u) \sim u^{-1}\phi(u)$ is quite sharp.

A.2. Multivariate Minimax Risk.

LEMMA A.3. *For any fixed n, s, r we have $B(r, \mathcal{M}(n, s_n)) = n\beta(s/n, r)$.*

PROOF. For any prior π on $\mathbb{R}^n$ and its marginals $\{\pi_i : i = 1, \ldots, n\}$ we have $\boldsymbol{\rho}(\boldsymbol{\theta}, \widehat{p}_{\pi_1\times\pi_2\times\ldots\pi_n}) = \sum_{i=1}^n \rho(\theta_i, \widehat{p}_{\pi_i})$. So,

$$B(\pi) = \int \pi(\theta)\boldsymbol{\rho}(\boldsymbol{\theta}, \widehat{p}_\pi)\, d\boldsymbol{\theta} \leq \int \pi(\boldsymbol{\theta})\boldsymbol{\rho}(\boldsymbol{\theta}, \widehat{p}_{\pi_1\times\pi_2\times\ldots\pi_n}) = \sum_{i=1}^n \int \pi(\theta_i)\rho(\theta_i, \widehat{p}_{\pi_i})\, d\theta_i$$

Again, if $\pi \in \mathcal{M}(n, s_n)$ then $\bar{\pi} = \pi_1\times\pi_2\times\ldots\pi_n \in \mathcal{M}_p(n, s_n)$ and $\mathcal{M}_p(n, s_n) \subset \mathcal{M}(n, s_n)$. So, $B(r, \mathcal{M}(n, s_n)) = B(r, \mathcal{M}_p(n, s_n))$ and due to decomposability of the Bayes risk for product priors we have

$$B(r, \mathcal{M}_p(n, s_n)) = \sup\left\{\sum_{i=1}^n \beta(\tau_i, r) : \sum_{i=1}^n \tau_i \leq s_n\right\}.$$

Now as $\beta(\tau, r)$ is concave function of τ the supremum in the above expression occurs when $\tau_i = s/n \,\forall\, i$. This completes the proof. □

Note that for any $\epsilon \in (0,1)$ the parametric space $\Theta(n, \epsilon\, s)$ as well as the prior space $\mathcal{M}(n, \epsilon\, s_n)$ is equivariant in the sense $\Theta(n, \epsilon s) = \epsilon \cdot \Theta(n, s)$ and $\mathcal{M}(n, \epsilon s_n) = \epsilon \cdot \mathcal{M}(n, s_n)$.

Proof of Lemma 5.2. From definition of Bayes risk it follows

$$\begin{aligned} B(\pi_{n,\epsilon}) &= \int \pi_{n,\epsilon}(\boldsymbol{\theta})\, \boldsymbol{\rho}(\boldsymbol{\theta}, \widehat{p}_{\pi_{n,\epsilon}})\, d\theta \le \int \pi_{n,\epsilon}(\boldsymbol{\theta})\, \boldsymbol{\rho}(\boldsymbol{\theta}, \widehat{p}_{\nu_{n,\epsilon}})\, d\boldsymbol{\theta} \\ &= \left\{ \int_{\Theta(n,s)} \nu_{n,\epsilon}(\boldsymbol{\theta}) \boldsymbol{\rho}(\boldsymbol{\theta}, \widehat{p}_{\nu_{n,\epsilon}})\, d\boldsymbol{\theta} \right\} \cdot \pi_{n,\epsilon}(\Theta_n) + \int_{\Theta^c(n,s)} \pi_{n,\epsilon}(\boldsymbol{\theta}) \boldsymbol{\rho}(\boldsymbol{\theta}, \widehat{p}_{\nu_{n,\epsilon}})\, d\boldsymbol{\theta} \\ &= \pi_{n,\epsilon}\big(\Theta(n,s)\big) B(\nu_{n,\epsilon}) + \int_{\theta \in \Theta^c(n,s)} \pi_{n,\epsilon}(\boldsymbol{\theta})\, \boldsymbol{\rho}(\boldsymbol{\theta}, \widehat{p}_{\nu_{n,\epsilon_n}})\, d\boldsymbol{\theta} \\ &\le \pi_{n,\epsilon}(\Theta(n,s)) R(n,s,r) + \mathrm{o}\left(B(r, \mathcal{M}(n, s_n))\right) \end{aligned}$$

as support of ν_n is contained in $\Theta(n,s)$, so we have $B(\nu_{n,\epsilon}) \le R(n,s,r)$ and we use property (c) on the second sum.

Now, using Condition (b) of the lemma, we have $R(n,s,r) \ge \epsilon\, B(r, \mathcal{M}(n, \epsilon s_n)) - \mathrm{o}\left(B(r, \mathcal{M}(n, s_n))\right)$ and the result follows by using the following Lemma A.4. □

Lemma A.4. *For any fixed $r \in (0, \infty)$ we have*

$$\lim_{\epsilon \uparrow 1} \liminf_{n \to \infty} \frac{B(r, \mathcal{M}(n, \epsilon s_n))}{B(r, \mathcal{M}(n, s_n))} = 1.$$

Proof. The proof is similar to Exercise 4.7 in Johnstone (2012). □

REFERENCES

Aitchison, J. (1975). Goodness of prediction fit. *Biometrika* **62** 547–554. MR0391353 (52 ##12174)

Aitchison, J. and Dunsmore, I. R. (1975). *Statistical prediction analysis*. Cambridge University Press, Cambridge. MR0408097 (53 ##11864)

Aslan, M. (2006). Asymptotically minimax Bayes predictive densities. *Ann. Statist.* **34** 2921–2938. . MR2329473 (2008g:62093)

Barndorff-Nielsen, O. E. and Cox, D. R. (1996). Prediction and asymptotics. *Bernoulli* **2** 319–340. . MR1440272 (97k:62006)

Barron, A. R. and Cover, T. M. (1988). A bound on the financial value of information. *IEEE Trans. Inform. Theory* **34** 1097–1100. . MR982823 (89k:90016)

Barron, A., Rissanen, J. and Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory* **44** 2743–2760. Information theory: 1948–1998. . MR1658898 (99h:94032)

Bell, R. M. and Cover, T. M. (1980). Competitive optimality of logarithmic investment. *Math. Oper. Res.* **5** 161–166. . MR571810 (81g:90114)

Brodiea, J., Daubechies, I., Molc, C. D., Giannoned, D. and Lorisc, I. (2009). Sparse and stable Markowitz portfolios. *Proceedings of National Academy of Sciences*.

Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist.* **42** 855–903. MR0286209 (44 ##3423)

Brown, L. (1974). Lecture Notes on Statistical Decision Theory. Available at "http://www-stat.wharton.upenn.edu/~lbrown".

Brown, L. D., George, E. I. and Xu, X. (2008). Admissible predictive density estimation. *Ann. Statist.* **36** 1156–1170. . MR2418653 (2009i:62023)

Brown, L. D. and Hwang, J. T. (1982). A unified admissibility proof. In *Statistical decision theory and related topics, III, Vol. 1 (West Lafayette, Ind., 1981)* 205–230. Academic Press, New York. MR705290 (84m:62013)

Buchdahl, J. (2003). *Fixed Odds Sports Betting: Statistical Forecasting and Risk Management*. High Stakes Publishing, London.

Candès, E. J. (2006). Compressive sampling. In *International Congress of Mathematicians. Vol. III* 1433–1452. Eur. Math. Soc., Zürich. MR2275736 (2008e:62033)

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. . MR2382644 (2009b:62016)

Cover, T. M. and Thomas, J. A. (1991). *Elements of information theory*. Wiley-Interscience, New York, NY, USA.

Csiszár, I. (1973). Generalized entropy and quantization problems. In *Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (Technical Univ. Prague, Prague, 1971; dedicated to the memory of Antonín Špaček)* 159–174. Academia, Prague. MR0359995 (50 ##12445)

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7** 269–281. MR520238 (80f:62016)

Dicker, L. H. (2012). Optimal Estimation and Prediction for Dense Signals in High-Dimensional Linear Models. arXiv:1203.4572.

Donoho, D. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture* 1–32.

Donoho, D. L. (2006). Compressed sensing. *IEEE Trans. Inform. Theory* **52** 1289–1306. . MR2241189 (2007e:94013)

Donoho, D. L. and Johnstone, I. M. (1994a). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. . MR1311089 (95m:62076)

Donoho, D. L. and Johnstone, I. M. (1994b). Minimax risk over l_p-balls for l_q-error. *Probab. Theory Related Fields* **99** 277–303. . MR1278886 (95g:62019)

Donoho, D., Johnstone, I. and Montanari, A. (2011). Accurate Prediction of Phase Transitions in Compressed Sensing via a Connection to Minimax Denoising. arXiv:1111.1041.

Donoho, D. L., Maleki, A. and Montanari, A. (2011). The noise-sensitivity phase transition in compressed sensing. *IEEE Trans. Inform. Theory* **57** 6920–6941. . MR2882271 (2012i:94068)

Donoho, D. L., Johnstone, I. M., Hoch, J. C. and Stern, A. S. (1992). Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B* **54** 41–81. With discussion and a reply by the authors. MR1157714 (93d:62010)

Efron, B. (2011). Tweedies Formula and Selection Bias. *Journal of the American Statistical Association* **106** 1602-1614.

Fan, J., Lv, J. and Qi, L. (2011). Sparse High-Dimensional Models in Economics. *Annual Review of Economics*.

FOSTER, D. P. and GEORGE, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22** 1947–1975. . MR1329177 (96c:62119)

FOURDRINIER, D., MARCHAND, É., RIGHI, A. and STRAWDERMAN, W. E. (2011). On improved predictive density estimation with parametric constraints. *Electron. J. Stat.* **5** 172–191. . MR2792550 (2012h:62088)

GEISSER, S. (1971). The inferential use of predictive distributions. In *Foundations of statistical inference (Proc. Sympos., Univ. Waterloo, Ont., 1970)* 456–469. Holt, Rinehart and Winston of Canada, Toronto, Ont. With comments by V. P. Godambe, I. J. Good, W. J. Hall and D. A. Sprott and a reply by the author. MR0381054 (52 ##1951)

GEISSER, S. (1993). *Predictive inference. Monographs on Statistics and Applied Probability* **55**. Chapman and Hall, New York. An introduction. MR1252174 (95k:62006)

GEORGE, E. I., LIANG, F. and XU, X. (2006). Improved minimax predictive densities under Kullback-Leibler loss. *Ann. Statist.* **34** 78–91. . MR2275235 (2008h:62034)

GEORGE, E. I., LIANG, F. and XU, X. (2012). From minimax shrinkage estimation to minimax shrinkage prediction. *Statist. Sci.* **27** 82–94. . MR2953497

GHOSH, M., MERGEL, V. and DATTA, G. S. (2008). Estimation, prediction and the Stein phenomenon under divergence loss. *J. Multivariate Anal.* **99** 1941–1961. . MR2466545 (2009m:62027)

HARTIGAN, J. A. (1998). The maximum likelihood prior. *Ann. Statist.* **26** 2083–2103. . MR1700222 (2000h:62030)

HUBER, N. and LEEB, H. (2012). Shrinkage estimators for prediction out-of-sample: Conditional performance. arXiv:1209.0899.

JOHNSTONE, I. M. (1994). On minimax estimation of a sparse normal mean vector. *Ann. Statist.* **22** 271–289. . MR1272083 (95g:62020)

JOHNSTONE, I. M. (2012). Gaussian Estimation: Sequence and Wavelet Models. Available at: "http://www-stat.stanford.edu/~imj".

KOMAKI, F. (1996). On asymptotic properties of predictive distributions. *Biometrika* **83** 299–313. . MR1439785 (98d:62048)

KOMAKI, F. (2001). A shrinkage predictive distribution for multivariate normal observables. *Biometrika* **88** 859–864. . MR1859415

KOMAKI, F. (2004). Simultaneous prediction of independent Poisson observables. *Ann. Statist.* **32** 1744–1769. . MR2089141 (2005k:62223)

KULLBACK, S. and LEIBLER, R. A. (1951). On information and sufficiency. *Ann. Math. Statistics* **22** 79–86. MR0039968 (12,623a)

LARIMORE, W. E. (1983). Predictive inference, sufficiency, entropy and an asymptotic likelihood principle. *Biometrika* **70** 175–181. . MR742987 (85k:62005)

LEEB, H. (2009). Conditional predictive inference post model selection. *Ann. Statist.* **37** 2838–2876. . MR2541449 (2011d:62145)

LIANG, F. (2002). *Exact minimax procedures for predictive density estimation and data compression*. ProQuest LLC, Ann Arbor, MI Thesis (Ph.D.)–Yale University. MR2703233

LIANG, F. and BARRON, A. (2004). Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Trans. Inform. Theory* **50** 2708–2726. . MR2096988 (2005f:94040)

LIANG, F. and BARRON, A. (2005). *Exact Minimax Predictive Density Estimation and MDL. Advances in Minimum Description Length: Theory and Applications (P. Grunwald, I. Myung and M. Pitt eds)*. MIT Press.

MAGEE, C. (2011). *Automatic Exchange Betting*. "http://www.betwise.co.uk".

MCMILLAN, B. (1956). Two inequalities implied by unique decipherability. *Information Theory, IRE Transactions on* **2** 115 -116.

Mukherjee, G. and Johnstone, I. M. (2012a). Minimax Risk of Predictive Density Estimation Over ℓ_p balls. Preprint.

Mukherjee, G. and Johnstone, I. M. (2012b). On the within-family Kullback-Leibler risk in Gaussian Predictive models. arXiv:1212.0325 [math.ST].

Murray, G. D. (1977). A note on the estimation of probability density functions. *Biometrika* **64** 150–152. MR0448690 (56 ##6995)

Ng, V. M. (1980). On the estimation of parametric density functions. *Biometrika* **67** 505–506. . MR581751 (81h:62077)

Nussbaum, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24** 2399–2430. . MR1425959 (98k:62065)

Pinsker, M. S. (1980). Optimal filtration of square-integrable signals in Gaussian Noise. *Problems in Information Transmission*.

Raskutti, G., Wainwright, M. and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q balls. *IEEE Transactions of Information Theory* **57** 6976–6994.

Rissanen, J. (1984). Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory* **30** 629–636. . MR755791 (85g:94009)

Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I* 157–163. University of California Press, Berkeley and Los Angeles. MR0084919 (18,947e)

Stein, C. (1974). Estimation of the mean of a multivariate normal distribution. In *Proceedings of the Prague Symposium on Asymptotic Statistics (Charles Univ., Prague, 1973), Vol. II* 345–381. Charles Univ., Prague. MR0381062 (52 ##1959)

Strawderman, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Statist.* **42** 385–388. MR0397939 (53 ##1794)

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of National Academy of Sciences*.

Vajda, I. (2002). On convergence of information contained in quantized observations. *IEEE Trans. Inform. Theory* **48** 2163–2172. . MR1930280 (2003j:94045)

Xu, X. and Liang, F. (2010). Asymptotic minimax risk of predictive density estimation for non-parametric regression. *Bernoulli* **16** 543–560. . MR2668914 (2011g:62110)

Xu, X. and Zhou, D. (2011). Empirical Bayes predictive densities for high-dimensional normal models. *J. Multivariate Analysis* **102** 1417-1428.

Zhang, C.-H. (2005). General empirical Bayes wavelet methods and exactly adaptive minimax estimation. *Ann. Statist.* **33** 54–100. . MR2157796 (2007a:62016)

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. . MR2604701 (2011d:62211)

Address of the First and Second authors
Department of Statistics
Sequoia Hall, 390 Serra Mall
Stanford University
Stanford, CA 94305-4065
E-mail: gourab@stanford.edu
imj@stanford.edu